APPENDIX D (DRAFT)

INTERVAL ESTIMATORS AND HYPOTHESIS TESTS FOR DATA QUALITY ASSESSMENTS IN WATER QUALITY ATTAINMENT STUDIES

Michael Riggs, Dept. Statistical Research, Research Triangle Institute

Elvessa Aragon, Dept. Statistical Research, Research Triangle Institute

Acknowledgements:

We are grateful to Andy Clayton (RTI), Florence Fulk (EPA/NERL), Laura Gabanski (EPA/OWOW), Susan Holdsworth (EPA/OWOW), Forrest John (EPA/Region 6), and Roy Whitmore (RTI) for providing thoughtful reviews of earlier drafts of this appendix and for suggesting a number of changes which greatly improved the final version.

Appendix D Table of Contents

Appendix D

Interval Estimators and Hypothesis Tests for Water Quality Attainment Studies

D.0 Introduction

This appendix provides formulae for the calculation of inferential statistics and sample size estimates necessary to complete the DQO and DQA processes for design and analysis of water quality attainment studies. These include confidence interval estimators for means, proportions, variances and percentiles, formulae for estimating minimum sample sizes necessary to compute confidence intervals of prespecified half-widths, and instructions for computing one-sample test statistics that can be used to test hypotheses about the attainment of water quality standards. Examples of the computation of all estimators and test statistics accompany the text. Special attention is paid to the exact binomial and exact hypergeometric tests.

Although all of the computations described in this appendix can be completed with statistical tables that are readily available in standard introductory texts (e.g., Lindley and Scott 1995; Steel et al. 1996) or on the Internet (e.g., http://www.epa.gov/quality1/qa_docs.html) and a hand calculator, in practice, they are usually done with the aid of statistical software. EPA provides software (DEFT) that may be used in the DQO process for sample size estimation for 1-sample t-tests and z-tests. EPA provides an additional software package (DataQuest) that can be used to carry out these tests and the large-sample Wilcoxon signed ranks test. Both packages can be downloaded from, http://www.epa.gov/quality1/qa_docs.html.

Although EPA does not endorse specific commercially available software, the following statistical packages are widely available at many academic and public research institutions. Perhaps the most comprehensive and widely used are SAS (http://www.sas.com/), SPLUS (http://www.insightful.com/products/default.asp), and its freeware counterpart R (http://www.r-project.org/). All of these packages provide well documented procedures and functions for a wide variety of statistical tests and estimators including those described in this appendix; however, they all require users to be conversant in their respective programming languages. By contrast, the SPLUS EnvironmentalStats package (http://www.probstatinfo.com/; Millard and Neerchal 2001), specifically designed to support the DQO and DQA processes, has an easy to use GUI interface through which all of the estimators, tests and sample size computations in this appendix may be produced with a few mouse clicks. Similarly, PASS (http://www.ncss.com/pass.html) and CIA (http://www.som.soton.ac.uk/cia/; Altman et al. 2000) are relatively inexpensive, menu-driven software for (respectively) power/sample size estimation and confidence interval estimation.

D.1. The Confidence Interval as a Tool for Specifying/Controlling Decision Error Rates

Formulae for constructing two-sided 100×(1-α)% confidence intervals from sample estimates of population means, binomial proportions, and variances are presented in Box 1a and are illustrated with sample calculations in Box 1b of this Appendix. These and all other confidence interval formulae in this section are extensions of the general formulae for 1- and 2-sided confidence intervals that were presented in Box 2 of Appendix C. General characteristics, utility and interpretation of confidence interval estimates are reviewed in Section C.1.5 of Appendix C. All of the 2-sided computations illustrated in this section can be easily extended to the case of 1-sided estimators on the basis of the relationships defined in Box 2 of Appendix C.

The expressions for confidence intervals on means and proportions (Box 1 of this Appendix) can be rearranged to solve for the sample size (n). Sample size formulae permit one to estimate *a priori* (i.e., during the DQO process) the minimum number of sampling units that must be collected in order to yield a confidence limit whose width is less than a specified maximum. For example the sample size required for $100\times(1-\alpha)\%$ two-sided confidence interval on the mean, with a half-width of W, can be obtained from the iterative solution of the first equation in Box 2a. Similarly, the sample size required for a $1-\alpha\%$ confidence interval on a binomial proportion with a half-width of W can be calculated with the second expression in Box 2a. Sample sizes for one-sided confidence intervals can be obtained by replacing the $1-\alpha/2$ t- and z-statistics in Box 2a with the corresponding $1-\alpha$ statistics. In either case, sample size calculations based on these formulae require the investigator to specify her desired maximum allowable Type I error rate, the standard error on the estimate and her desired maximum half width for the confidence interval. Optimally, the standard error should be obtained from either a pilot study or some previous study on the same target population (e.g., a previous assessment); failing that, the investigator may choose to use a standard error from a published study of a similar population. When α and SE are fixed, specifying the maximum acceptable confidence interval half-width is comparable to controlling the Type II error rate in a statistical hypothesis test.

Consider the case of an investigator who desired 95% confidence limits on the acute CMC of selenite from a sample of n 1-liter volumes taken from a stream reach, such that the confidence interval half-width (W) would be ≤ 20 µg/L. A pilot study indicated that the standard deviation for selenite concentration in the stream was 200. Applying the first equation in Box 2a, he determined that he needed to collect 387 1-liter volumes from the stream reach to meet this requirement. Alarmed, he recomputed n, this time specifying w=50 µg/L and was relieved find the new sample size requirement of 64 volumes. This demonstrates the point made earlier; simultaneous control of α and β to low levels requires extremely large sample sizes. The only way he could maintain an α of 0.05, without increasing the sample size, was to increase the confidence interval width, thereby increasing the Type II error rate and decreasing the precision of his estimate. However, if he is willing to accept a higher Type I error rate, say $\alpha=0.15$, he could maintain a half-width of $\leq$ 20 µg/L on an 85% confidence interval with a sample size of only n=209 1-liter volumes, as compared to n=387 for a 95% confidence interval with the same half-width. Detailed discussion of the complex interrelationships among specified -levels and half-widths, variances and other factors that affect confidence intervals and hypothesis tests are taken up in Sections C.3.1 and C.3.2 of Appendix C.

Box 1-a : Computation of 100 x (1-a)% Confidence Intervals for Sample Estimates

Population Mean

Suppose that $c_1, c_2..., c_n$ represents a random sample of $n \geq 30$ data points from an essentially infinite population of sampling units.

Step 1. Calculate the sample mean $\bar{c}$ and the sample variance s^2 (see Appendix C, Box 1-a and 1-b).

Step 2. Use the table of percentage points of student's t-distribution to determine the value $t_{1\text{-}a/2,n\text{-}1}$ such that 100 x (1-a/2)% of the student t-distribution with n-1 degrees of freedom is below $t_{1\text{-}a/2,n\text{-}1}$.

Step 3. The 100 x (1-a)% confidence interval for the population mean is bounded by the following upper and lower limits:

$$\bar{c} - t_{1-a/2,\, n-1}\sqrt{\frac{s^2}{n}} \text{ to } \bar{x} + t_{1-a/2,\, n-1}\sqrt{\frac{s^2}{n}} .$$

Step 4. If n is ≥ 60, the value of $t_{1-a/2,n-1}$ may be replaced by the value $z_{1-a/2}$ of the standard normal distribution (the table of percentage points of student's t-distribution).

Population Binomial Proportion

Suppose that c is a dichotomous random variable with values 0, 1 (denoting, for example, the presence of absence of some observable characteristic). Let $c_1, c_2..., c_n$ represent a random sample of n data points from a population of c values. If n is ≥ 50, an approximate $100 \times (1 - a/2)\%$ confidence interval can be computed as follows:

Step 1. Calculate the sample proportion p: $p = \frac{1}{n}\sum_{i=1}^{n} c_i$

Step 2. Use the table of the cumulative probabilities of the standard normal distribution to determine the value of $z_{1-a/2}$ such that $100 \times (1 - a/2)\%$ of the standard normal distribution is below $z_{1-a/2}$.

Step 3. The large-sample $100 \times (1 - a/2)\%$ confidence interval for the population proportion is bounded by the following lower and upper limits:

$$p - z_{1-a/2}\sqrt{\frac{p(1-p)}{n}} \text{ to } p + z_{1-a/2}\sqrt{\frac{p(1-p)}{n}} .$$

Population Variance

Suppose that $c_1, c_2..., c_n$ represents a random sample of n data points from a population of normally distributed values for c .

Step 1. Calculate the sample variance s^2 (see Appendix C, Box 1-a).

Step 2. Use Table of critical values of the chi-square distribution to determine the value of $c^2_{1-a/2,\, n-1}$ such that) $100 \times (1 - a/2)\%$ of the chi-square distribution with n-1 degrees of freedom below $c^2_{1-a/2,\, n-1}$. Similarly, determine the value $c^2_{a/2,\, n-1}$ such $100 \times (a/2)\%$ of the chi-square distribution with n-1 degrees of freedom is below $c^2_{a/2,\, n-1}$.

Box 1-a : (Continued) Calculating of 100 x (1-a)% Confidence Intervals

Step 3. The 100 x (1-a)% confidence interval for the population mean is bounded by the following upper and lower limits:

$$\frac{(n-1)s^2}{c^2_{1-a/2,\,n-1}} \;,\; \frac{(n-1)s^2}{c^2_{a/2,\,n-1}} .$$

Note :
The **width** of the confidence interval is the distance between the upper and lower confidence limits. The confidence intervals for the mean and proportion are symmetric; that is, the distance between the estimates and their upper and lower bounds (**half-width**) are equal. The confidence interval for the variance is asymmetric.

Box 1-b : Computation of 100 x (1-a)% Confidence Intervals for Sample Estimates

Consider the random sample of 10 turbidity measurements from Appendix C, Box 1-b. Suppose that the population of turbidity measurements is normally distributed. (Later sections will discuss how to determine if a normal distribution may be assumed for a population.) To calculate 95% confidence intervals for the mean, proportion and variance, follow the steps outlined in Box 1-a of this Appendix.

Population Mean

Step 1. The sample mean and sample variance were calculated in Box 1-b:

$$c = 101.3 \quad \text{and} \quad s^2 = 8111.7889$$

Step 2. Since $df = 10 - 1 = 9$, and $1\text{-}a = 0.95$, $a = 0.05$, $1 - a/2 = 1 - 0.05/2 = .0975$, use the Table of percentage points of student's t-distribution to determine the value $t_{0.975,9}$ of the distribution with 9df. This value is 2.262.

Step 3. Calculate the lower and upper limits.

$$101.3 - 2.262\sqrt{\frac{8111.7889}{10}} = 101.3 - 64.42 = 36.88$$

$$101.3 + 2.262\sqrt{\frac{8111.7889}{10}} = 101.3 + 64.42 = 165.72$$

The 95% confidence interval for the population mean is (36.88, 165.72). Note that the width of the confidence interval is 64.42.

Population Proportion

Note that the sample size *n=10* is not large enough to use the steps outlined in Box 1-a of this Appendix. The exact binomial confidence interval discussed in a later section will appropriate for small sample sizes.

For purposes of showing a sample calculation, suppose that we have a sample of 75 turbidity measurements from the same population.

Step 1. Suppose that the sample proportion was calculated to be *p=0.18*.

Step 2. Use the Table of the cumulative probabilities of the standard normal distribution to determine the value $z_{0.975}$ of the standard normal distribution. This value is 1.96.

Step 3. The lower and upper limits are

$$0.18 - 1.96\sqrt{\frac{0.18(1-0.18)}{10}} = .018 - 0.24 = -0.06$$

$$0.18 + 1.96\sqrt{\frac{0.18(1-0.18)}{10}} = .018 + 0.24 = 0.42$$

Note that the lower limit is -0.06. Since proportions cannot be negative, the lower limit becomes 0. The 95% confidence interval for the population proportion is (0, 0.42). Note that the width of the confidence interval is 0.24.

Box 1-b: (Continued) Examples for Calculating of 100 x (1-a)% Confidence Intervals

Population Variance

Step 1. From Box 1-b of Appendix C, s^2=8111.8.

Step 2. Use the Table of critical values of the chi-square distribution to find the values $c^2_{0.975,9}$ and $c^2_{0.025,9}$ of the chi-squared distribution with 9 df. These values are 19.023 and 2.700, respectively.

Step 3. The lower and upper 95% confidence limits on the sample estimate of the variance (s^2) of the turbidity values in Mermentau River, June 1980 - April 2000, are

$$\frac{(10-1)8111.8}{19.02} = 3837.8 \quad \text{and} \quad \frac{(10-1)8111.8}{2.70} = 27039.3$$

The 95% confidence interval for the population variance is (3837.7806, 27039.2963). Note that the confidence interval is asymmetric; the lower confidence interval has a width of 4274 (8111.8-3837.8), while the width of the upper confidence interval is 18,927.5 (27039-8111.8). This reflects the asymmetry of the chi-square distribution which is used to model the variance. Note also the extreme width of the overall confidence interval (23,201.5). This demonstrates a general attribute of the population variance: unless s^2 is small, it is very difficult to obtain good estimates of s^2 from small samples. The population variances of most ambient environmental parameters (e.g. turbidity) and natural population parameters will usually be quite large.

Box 2-a : Minimum Sample Size Requirements for Estimating a Population Mean of Binomial Proportion with 100 x (1-a)% Confidence.

Population Mean

Step 1a. Decide on the level of confidence (1-a) and maximum acceptable half-width (W) of the 100 x (1-a)% confidence that you will be calculating.

Step 2a. Iteratively search for the smallest value of n that satisfies the equality:

$$n = \left(t_{1-a/2,\, n-1} c \, \frac{s}{W} \right)^2 \qquad \text{where } s = \sqrt{s^2}$$

Step 3a. Typically, *n* will not be an integer. Round up the calculated value of *n* to the next integer to obtain the minimum sample size required to insure a 100 x (1-a)% confidence interval whose half-width is no larger than W.

Population Proportion

Step 1b. Decide on the level of confidence (1-a) and maximum acceptable half-width (W) of the 100 x (1-a)% confidence interval you will be calculating.

Step 2b. Solve the following algebraic expression for *n*:

$$n = \frac{1}{\left(\sqrt{z^2_{1-a/2}\, p(1-p) + 2W} - z_{1-a/2}\sqrt{p(1-p)} \right)^2}$$

Step 3b. Round up the calculated value of *n* to the next integer to obtain the minimum sample size required to insure a 100 x (1-a)% confidence interval whose half-width is no larger than W.

Note:

While the minimum sample size "n" appears on both sides of the equation used to get *n* for the mean (2a of this Appendix), it appears only on the left side of the equation that is used to get a minimum *n* for proportion (2b of this Appendix). The practical consequence of this is that we can use basic algebra to solve for the *n* needed to construct a confidence interval for a proportion, but we must use a trial-and-error approach (i.e. iteration) to get the corresponding minimum *n* for the case of the mean. Ideally, one should write a computer program with a "do-loop" to generate a large number of candidate solutions for *n*, computed over a wide range of df values (i.e., n-1) for the t-statistic on the left hand side of the equation in step 2a. However, for those without access to a computer, we provide in Box 2-b of this Appendix, a more tedious solution that can be done with the aid of a hand calculator.

The confidence interval formulae in Box 1a require the assumption that repeated sample estimates of the population parameters be normally distributed about the true population parameter. This assumption has been proven to hold for a large number of sample estimators that are based on an estimation procedure called **maximum likelihood**. In general, the normality assumption will hold for the sample means and proportions based on n > 20-30. When sample sizes are <20, different estimation methods should be considered. For example, confidence intervals for binomial proportions from samples with sample sizes (n) such that $n \times p = 5.0$ and/or $n \times (1-p) = 5.0$, should be computed using exact binomial methods. Exact confidence intervals and associated minimum sample size requirements can be obtained from tables (e.g., CRC Basic Statistical Tables) or from widely available statistical software packages such as PASS, SPLUS EnvironmentalStats, and SAS (O'Brien 1998; http://www.bio.ri.ccf.org/power.html).

Two-sided confidence intervals for the sample median or for sample percentiles (e.g., the 95th percentile) can be computed using nonparametric methods; so-called because they don't assume a particular underlying parametric distribution such as the normal. Like the above-described methods, the nonparametric methods have both large-sample (n>20) and small sample exact forms. Nonparametric exact confidence intervals and minimum sample size requirements can be obtained from tables in nonparametric statistics books (e.g., Hollander and Wolfe 1999) or from statistical software packages [e.g., SAS (PROC FREQ and PROC NPAR1WAY)]. The large sample formulae for $1-\alpha\%$ confidence intervals on the population median or on any of its percentiles are based on the ranks of the data in the sample. The ranks are obtained by first sorting the data from smallest to largest, finding the value which corresponds to the median or desired percentile and then computing the ranks of the observations which correspond to the lower (r) and upper (s) $1-\alpha\%$ bounds on the sample estimate. The formulae for computing the median and other population percentiles and r and s for their confidence intervals are shown in Boxes 3 and 4, respectively.

For example, consider the 244 monthly turbidity values in Table 3 of Appendix C. The data are sorted from smallest to largest beginning with the first value in row one and increasing from left to right within the row. The largest values are in the last row (row 49). Because the sample size (244) is an even number, the median is computed as the mean of the middle two observations (i.e., 75 and 76, bold in row 24). To compute the 2-sided 90% confidence interval on the median, we apply the first two equations in Box 4. First, we find Z associated with 1-(0.10/2), which is 1.645. Next, we solve the equations to find that the ranks of the 90 % lower and upper bounds on the median turbidity correspond to 109.15 and 135.8. However, ranks must be integers, so r needs to be rounded *down* to the nearest integer (109) and s needs to be rounded *up* to the nearest integer (136). The turbidity values corresponding to these ranks (bold in Appendix C Table 3) are 70 and 80; thus:

2-sided 90% confidence interval on the median = 75.5 (70.0, 80.)

In addition, using the equations in Box 5a, we can easily write the 1-sided 90% confidence intervals for the median turbidity:

Box 3-a : Sample Percentiles and the Sample Median

In a population or sample, a percentile is the data value that is greater than or equal to a given percentage of all the data values. For instance, the 90th percentile is the data value that is greater than or equal to 90% of all the data values. The 50th percentile is more commonly called the sample median.

Let $c_1, c_2..., c_n$ represent a random sample of *n* population units, ordered from the smallest to the largest values of c_i. The pth percentile, is the value of c associated with the $c(p)^{th}$ ordered value from the sample size *n*; i.e. $c(p)$ is the rank of c value which is $\geq$ p-percent of the other values in the sample. $c(p)$ can be determined by the following simple 4-step process:

Step 1. Order the data values from smallest to highest and label these ordered values $c_{(1)}, c_{(2)}..., c_{(n)}$. Thus $c_{(1)}$ is the smallest value, $c_{(2)}$ is the second smallest, and $c_{(n)}$ is the largest.

Step 2. Calculate np/100. Separate the integer part of the result. That is,

$$\frac{np}{100} = i + f, \qquad \text{where } i \text{ is the integer part and } j \text{ is the fraction part}.$$

For instance, 135.78=135+0.78.

Step 3. If *f=0*, then $c(p) = \dfrac{c_{(i)} + c_{(i+1)}}{2}$. Otherwise, $c(p) = c_{(i+1)}$.

Step 4. For the sample median, *p=50.* Hence, *n*(50)/100=*n*/2; i.e. the value of c associated with $n/2^{th}$ ordered sample value is the sample median.

Box 3-b : Example for Calculating Sample Percentiles and the Sample Median

Consider the 10 turbidity measurements from Box 1-b in Appendix C : 34, 58, 87, 95145, 14, 38, 62, 95, 160, 320.

Step 1. The ordered values are: 14, 34, 38, 58,62, 87, 95, 145, 160, 320.
Step 2. Calculate the 25th, 50th, 75th and 90th percentiles.

$$p = 25 : \frac{(10)(25)}{100} = 2.5 = 2 + 0.5$$

$$p = 50 : \frac{(10)(50)}{100} = 5 = 5 + 0$$

$$p = 75 : \frac{(10)(75)}{100} = 7.5 = 7 + 0.5$$

$$p = 90 : \frac{(10)(90)}{100} = 9 = 9 + 0$$

Step 3. Determine y(p).

p	i	f	$y(p)$
25	2	0.5	$\frac{c_{(2)} + c_{(3)}}{2} = \frac{(14 + 34)}{2} = 2$
50	5	0	$c_{(5)} = 62$
75	7	0.5	$\frac{c_{(7)} + c_{(8)}}{2} = 95 + 145/2 = 120$
90	9	0	$c_{(10)} = 320$

Box 4-a : Large-Sample 100 x (1-a)% Confidence Intervals for Percentiles

Population Median

Let $c_{(1)}, c_{(2)} ..., c_{(n)}$ represent the ordered values in a random sample of *n* units, such that ≤ 20 from a population. Calculate

$$r = \frac{n}{2} - \left(z_{1-a/2} c \frac{\sqrt{n}}{2} \right)$$

$$s = 1 + \frac{n}{2} + \left(z_{1-a/2} c \frac{\sqrt{n}}{2} \right)$$

Round *r* down to the next lower integer and round *s* up to the next highest integer. Find the rth and sth ordered values in the sample. The large-sample 100 x (1-a)% confidence interval for the median is $\left(c_{(r)}, c_{(s)} \right)$.

pth Population Percentile

Let $c_{(1)}, c_{(2)} ..., c_{(n)}$ represent the ordered values in a random sample of *n* units from a population. Calculate

$$r = \frac{np}{100} - \left(z_{1-a/2} c \frac{\sqrt{np(100-p)}}{100} \right)$$

$$s = 1 + \frac{np}{100} + \left(z_{1-a/2} c \frac{\sqrt{np(100-p)}}{100} \right)$$

Round *r* down to the next lower integer and round *s* up to the next higher integer. Find the rth and sth ordered values in the sample. The large-sample 100 x (1-a)% confidence interval for the pth percentile is $\left(c_{(r)}, c_{(s)} \right)$.

Box 4-b : Examples for Calculating Large-Sample 100 x (1-a)% Confidence Intervals for Percentiles

Consider a random sample of 25 turbidity measurements from the Louisiana River between June 1980 and April 2000. The ordered (upper left to lower right) sample values are:

14	26	34	38	40
45	50	50	58	62
70	75	76	80	87
95	115	125	130	145
160	175	210	245	320

Calculate 95% confidence intervals for the median and the 90th percentile.

From the table of the cumulative probabilities of the standard normal distribution, $z_{0.975} = 1.96$.

For the median:
Using the steps in Box 3-a, the sample median is determined to be the 13thordered value, which is 76.

$$r = \frac{25}{2} - \left(1.96c\,\frac{\sqrt{25}}{2}\right) = 12.5 - 4.9 = 7.6 \quad \text{rounded down to 7}$$

$$s = 1 + \frac{25}{2} + \left(1.96c\,\frac{\sqrt{25}}{2}\right) = 13.5 + 4.9 = 18.4 \quad \text{rounded up to 19}$$

The 7th ordered value is 50, and the 19th ordered value is 130. Hence, the 95% confidence interval for the median is (50, 130).

For the 90th percentile:

Using the steps in Box 3-a, the sample 90th percentile is determined to be $\left(c_{(22)} + c_{(23)}\right)/2 = (175 + 210)/2 = 192.5$.

$$r = \frac{25}{2} - \left(1.96c\,\frac{\sqrt{25}}{2}\right) = 12.5 - 4.9 = 7.6 \quad \text{rounded down to 7}$$

$$s = 1 + \frac{25}{2} + \left(1.96c\,\frac{\sqrt{25}}{2}\right) = 13.5 + 4.9 = 18.4 \quad \text{rounded up to 19}$$

The 19th ordered value is 130. Since *n*=25, there is no 27th ordered value, so the largest ordered value, 320, becomes the upper limit. Hence, the 95% confidence interval for the 90th percentile is (130, 320).

Lower 1-sided 90% confidence interval on the median = 75.5 (70.0, $+\infty$)
Upper 1-sided 90% confidence interval on the median = 75.5 ($-\infty$, 80.0)

Because the median and the percentiles are based on ranks, the minimum sample size requirements for their estimation needs to be stated a little differently than for means, proportions or variances. Applying results from the theory of order statistics, we can calculate the minimum number of sampling units that is required for a sample to contain at least one value $\geq$ the value of the population median (second equation in Box 5a) or, if we desire, of the population p^{th} percentile (first equation in Box 5a), with 1-α% confidence. For example, if an investigator desires to have 95% confidence that a sample of sampling units from a lagoon contains at least one sampling unit whose selenium concentration is as large or larger than the 95th percentile of selenium concentration in the universe of all possible sampling units from the lagoon, the first equation in Box 6a dictates that his sample must contain at least 59 sampling units.

A final point to consider when using nonparametric methods is that the investigator does not have as much control over the Type II error rates as he does with the parametric methods. This is because he cannot specify a minimum confidence interval half-width; all he can control is the α-level, which he does by specifying a 1-α% confidence level. However, it is possible to obtain minimum sample size estimates for the median, which control for both α and β by applying methods based on the Wilcoxon Signed-Rank Test (both exact and large-sample versions; Hollander and Wolfe 1999).

The methods described to this point allow the specification of acceptable decision error rates based on specification of the 1-α% confidence level and, conditional on such specification, fixing the half-width of the confidence interval (and thereby the Type II error rate) by specifying a minimum sample size. For a given 1-α% confidence level and a standard value C, against which the sample estimate is to be compared, specifying (for example) a maximum allowable half-width of 0.10 x C will provide substantially more statistical power (i.e., 1-β) than will a specified half-width of 0.50 x C.

D.2 Parametric one-sample tests on means

The use of the one-sample t-test to evaluate attainment vs. impaired hypotheses involving the mean of a continuous water quality variable (e.g., turbidity) against a criterion value (e.g., the maximum allowable turbidity value in a stream) was discussed in detail in Sections C.2.1 Appendix C. The assumptions required by the test were listed in section C.3.1 and graphical methods for their verification were illustrated in Section C.2.1. In this section we briefly review the features and assumptions of the t-test and then focus on the problem of employing the test for analysis of data that are highly autocorrelated and confounded by significant seasonal effects, two commonly encountered problems in water quality studies.

The general form of the t-statistic is provided in Box 6. When the null hypothesis is true and the sampling units have been independently sampled from a population in which the attribute values are normally distributed, t will follow a t-distribution with n-1 degrees of freedom (df). But if the alternative hypothesis is true, it will follow a noncentral t-distribution with df=n-1.

Box 5-a : Minimum Sample Size Requirements for Constructing 100 x (1-a)% Confidence Intervals for Percentiles

If we want to have to 100 x (1-a)% confidence that a random sample of size *n* from a target population contains a one value of c (e.g., turbidity) that is at least as large as the pth population percentile of c , the minimum sample size can be computed as:

$$n = \frac{\ln\left[1-(1-a)\right]}{\ln\left(\frac{p}{100}\right)}$$

For the median (p=50), this becomes

$$n = \frac{\ln\left[1-(1-a)\right]}{}$$

Box 5-b : Examples for Calculating the Minimum Sample Size Requirements for Constructing 100 x (1-a)% Confidence Intervals for Percentiles

Suppose we are in a monitoring situation wherein we need to be 95% certain that we regularly obtain samples from the sediment that contain at least 1 sediment aliquot (i.e., sampling unit) whose concentration of selenium is at or above the median of selenium concentrations in the reservoir that is being monitored? At or above the 90th percentile?

Question 1: How many sample units (i.e., aliquots of sediment) must the sample contain in order to be 95% confident that the sample contains at least one aliquot whose selenium concentration is as large as, or larger than, the median (=50th percentile) of the selenium concentration of the entire target population?

Letting p=0.50 and $a = 0.01$ in the equation in Box 3-a, *n* is computed as:

$$n = \frac{\ln(1-0.95)}{\ln(0.50)} = \frac{\ln(0.05)}{\ln(0.50)} = 4.32 \qquad \text{rounded up to 5}$$

Question 2: Similarly, how many sampling units must the sample contain in order to be 95% confident that the sample contains at least one aliquot whose selenium concentration is as large as, or larger than, the 90th percentile of the selenium concentration of the entire target population?

Letting p=0.90 and $a = 0.01$ in the equation in Box 3-a, *n* is computed as:

$$n = \frac{\ln(1-0.95)}{\ln(0.90)} = \frac{\ln(0.05)}{\ln(0.90)} = 28.43 \qquad \text{rounded up to 29}$$

Box 6-a: One-Sample t-Test for the Mean
(One-sided Case)

Assumptions

1. The distribution of the attribute X is normal with population mean *μ* and variance s^2.
2. X_1, X_2, …, X_n is an independent sample of *n* individuals from the target population.

Let μ_0 be the fixed criterion against which the population mean μ is compared. Consider each of the hypothesis pairs:

Case 1: $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$
Case 2: $H_0 : \mu = \mu_0$ vs. $H_a : \mu < \mu_0$

For either case, when the two assumptions hold, a one-sample t-test for the null hypothesis vs. the alternative hypothesis can be performed as follows:

Step 1. Select the significance level *a*. (Typical values of *a* are 0.05 and 0.01.)
Step 2. Calculate the sample mean $\bar{x}$ and the sample variance s^2 (see Box 1-a).
Step 3. Use the table of percentage points of student's t-distribution to find the value $t_{1-a,\ n-1}$ such that (1-a)x100% of the Student t distribution with n-1 degrees of freedom is below $t_{1-a,\ n-1}$.
Step 4. Calculate the test statistic:

$$t_c = \frac{\bar{x} - m_0}{\sqrt{\frac{s^2}{n}}}$$

where $\bar{x}$ = the sample mean of the measured attribute, X
s^2 = the sample variance of the measured attribute, X
n = the number of sampling units in the sample
μ_0 = the fixed criterion against which the population mean is compared.

Step 5. Compare t_c with $t_{1-a,\ n-1}$.
Case 1: If $t_c > t_{1-a,\ n-1}$ then reject H_0.
If $t_c = t_{1-a,\ n-1}$ then reject H_0.
Case 2: If $t_c < t_{1-a,\ n-1}$ then reject H_0.
If $t_c = t_{1-a,\ n-1}$ then reject H_0.

Box 6-b: Example for Performing a One-Sample t-Test for the Mean

Consider a random sample of 35 turbidity measurements from the Mermentau River between June 1980 and April 2000. The sample values are:

34	58	80	87	145	245	26
40	50	75	130	175	14	38
62	76	95	160	45	115	210
320	50	70	125	8	432	52
26	19	32	80	85	170	352

It is desired to test the null hypothesis that the mean monthly turbidity measurement is no more than 150 NTU vs. the alternative the mean is greater than 150 NTU.

$$H_0 : \mu \leq 100 \quad \text{vs.} \quad H_a : \mu > 100$$

Step 1. The desired significance level is a=0.05.

Step 2. Using the equations in Box 1-b of Appendix C, the sample mean and sample variance are calculated to be $\bar{x}$ = 108.1 and s^2 = 9924.70.

Step 3. With a=0.05, 1-a/2=1-0.025=0.975 and df=35-1=34. The table of percentage points of student's t-distribution yields $t_{0.975,\ 34}$=2.032.

Step 4. Calculate the test statistic.

$$t = \frac{108.1 - 150}{\sqrt{\frac{9924.70}{35}}} = \frac{-41.9}{16.84} = -2.49$$

Step 5 Since -2.49 < 2.032, the null hypothesis cannot be rejected. Accept the null hypothesis and conclude that the true mean monthly turbidity seems to be less than or equal to 150 NTU.

While many things in nature tend to be normally distributed (e.g., weights of organisms), environmental data are very commonly log-normally distributed. Thus, *before* computing a t-test, it is important that the normality of the distribution of the sampling units be assessed using the methods described in Sections C.2.1 and C.2.2. If the data are approximately normal they can be analyzed as is; if not, then a suitable transformation (e.g. log-transformation) should be applied to the data before carrying out the t-test. Normality of the transformed variable can be verified using Q-Q plots as illustrated in C.2.2.

Frequently, good sampling and/or experimental design will be sufficient to insure independence among the sampling units. However, as pointed out in Section C.2.5, inherent temporal autocorrelation may be so strong that it will not be possible to design it away without substantial loss of data. We can illustrate the problem of strong autocorrelation and one of its possible solutions with some of the Mermentau River turbidity data. Recall that the turbidity data are approximately lognormal and autocorrelated for lags other than 3, 8 or 9 months. Suppose that we desire to test the null hypothesis that the mean turbidity during the most recent 3-year period (i.e., 1997-1999, inclusive) is less than or equal to the 150 NTU criterion vs. the 1-sided alternative that it is greater than 150 NTU. This requires that we consider the log-transformed values of the 36 monthly turbidity measurements that are listed in the fourth column of Table 1.

The first step is to plot the turbidity time series and the correlogram of the log-turbidity, as was done in the example in Section C.2.5 of Appendix C. These two plots (Figs. 1A and 1B) confirm the existence of both seasonality and significant autocorrelation in the 3-year time series. Not surprisingly, the patterns are quite similar to those of the 20-year series (Appendix C, Fig. 11). Significant autocorrelations occur between measurements that were taken 1, 2, 6, 7, 8, 18 and 19 months apart. The repeating annual pattern indicates strong seasonality in the series. In fact, the assessment of autocorrelation is difficult to make in the face of either seasonality or long-term trends. Thus it is desirable to remove their effects from the series before making a final of interpretation of the correlogram.

Seasonality and long-term trend are defined as fixed effect deviations from the mean of a time-series and can be estimated and removed from the series by the fitting of a two-way **ANOVA model**:

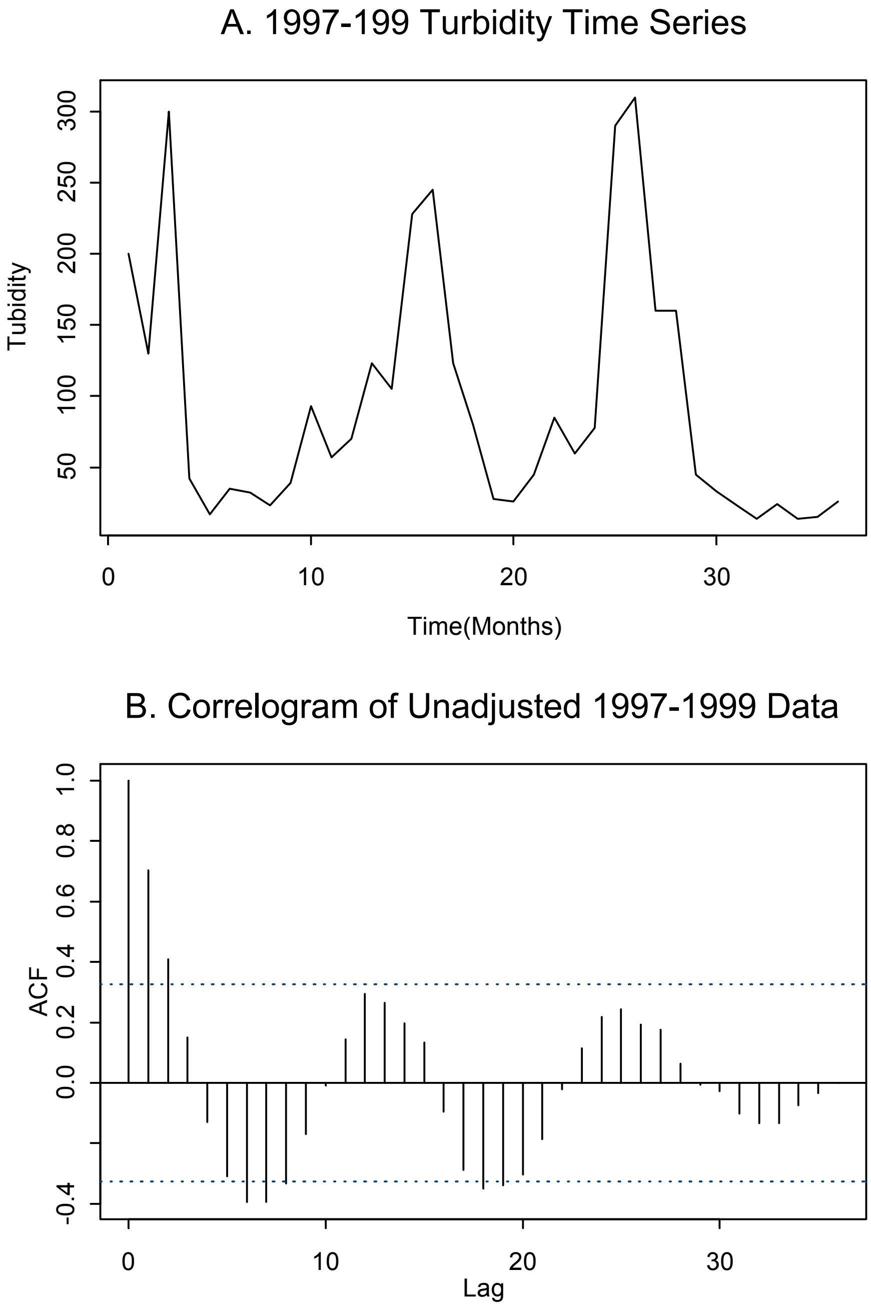
$$y_{ij} = m_0 + b_i Year_i + b_j Month_j + e_{ij} \qquad (1)$$

This model says that each observed monthly log-turbidity value (y_{ij}) is due to the 3-year mean of the time series (μ_0) + an effect due to the particular year in which it was taken (e.g., a dry year or a wet year) + the effect of the month of the year in which it was taken + random error. If there had been increasing development of the lands adjacent to the Mermentau River over the 1997-1999 period this may well have led to increasing erosion and runoff with corresponding annual increases in turbidity. If this had been the case, the year effect would have estimated the 3-year trend. Similarly, month-to-month differences that are consistent across years are an indication of seasonality and are estimated in the ANOVA model by the month effects. A cursory examination of the raw data (Table 1, column 3) suggests the presence of seasonality but not trend. The final term in the model, represented by the Greek epsilon (ε_{ij}), is what is left over after removing trend and seasonality from the overall 3-year average of the log-turbidity measurements. Thus, the ε_{ij} , called residuals, are the seasonally adjusted values that we desire.

Table 1. Intermediate calculations for a seasonally adjusted ttest on 1997-1999 Turbidity data.

				Adjusted	Squared	Adjusted
		Turbidity	Log of	Log of	Adjusted	x Lag1 Adj.
YEAR	MONTH	(NTU)	Turbidity	Turbidity	Values	Values
1997	1	200	5.298	0.070	0.005	.
	2	130	4.868	-0.187	0.035	-0.013
	3	300	5.704	0.332	0.110	-0.062
	4	42	3.738	-1.002	1.005	-0.333
	5	17	2.833	-0.953	0.908	0.955
	6	35	3.555	-0.225	0.050	0.214
	7	32	3.466	0.186	0.035	-0.042
	8	23	3.135	0.156	0.024	0.029
	9	39	3.664	0.146	0.021	0.023
	10	93	4.533	0.693	0.480	0.101
	11	57	4.043	0.459	0.211	0.318
	12	70	4.248	0.325	0.106	0.149
1998	1	123	4.812	-0.733	0.537	-0.239
	2	105	4.654	-0.717	0.514	0.526
	3	228	5.429	-0.258	0.067	0.185
	4	245	5.501	0.445	0.198	-0.115
	5	123	4.812	0.710	0.504	0.316
	6	80	4.382	0.286	0.082	0.203
	7	28	3.332	-0.264	0.070	-0.075
	8	26	3.258	-0.038	0.001	0.010
	9	45	3.807	-0.028	0.001	0.001
	10	85	4.443	0.286	0.082	-0.008
	11	60	4.094	0.194	0.038	0.056
	12	78	4.357	0.117	0.014	0.023
1999	1	290	5.670	0.663	0.440	0.078
	2	310	5.737	0.904	0.818	0.600
	3	160	5.075	-0.074	0.005	-0.067
	4	160	5.075	0.557	0.311	-0.041
	5	45	3.807	0.243	0.059	0.135
	6	33	3.497	-0.061	0.004	-0.015
	7	23	3.135	0.078	0.006	-0.005
	8	14	2.639	-0.118	0.014	-0.009
	9	24	3.178	-0.118	0.014	0.014
	10	14	2.639	-0.979	0.958	0.115
	11	15	2.708	-0.653	0.427	0.640
	12	26	3.258	-0.443	0.196	0.289
		Totals	148.400		8.348	3.955

Fig. 1. (A) Time series plots of 36 monthly turbidity measurements from the Mermantau River, January 1997- December 1999. (B) Correlograms showing seasonality and autocorrelation of log-transformed turbidity measurements.


A. 1997-199 Turbidity Time Series

B. Correlogram of Unadjusted 1997-1999 Data

The simplest way to obtain the residuals is to fit the two-way ANOVA model with a statistical software package (e.g., SAS, SPLUS, etc.). These software packages will automatically compute the residuals and save them in a data set, which the analyst can use for the remaining calculations in this section. The values in column 5 of Table 1 (Adjusted Log of Turbidity) are the residuals from the fitted two-way ANOVA model of the log-turbidity data.

Alternatively, if such software is not available, the analyst can easily do the necessary computations with a hand calculator, as follows:

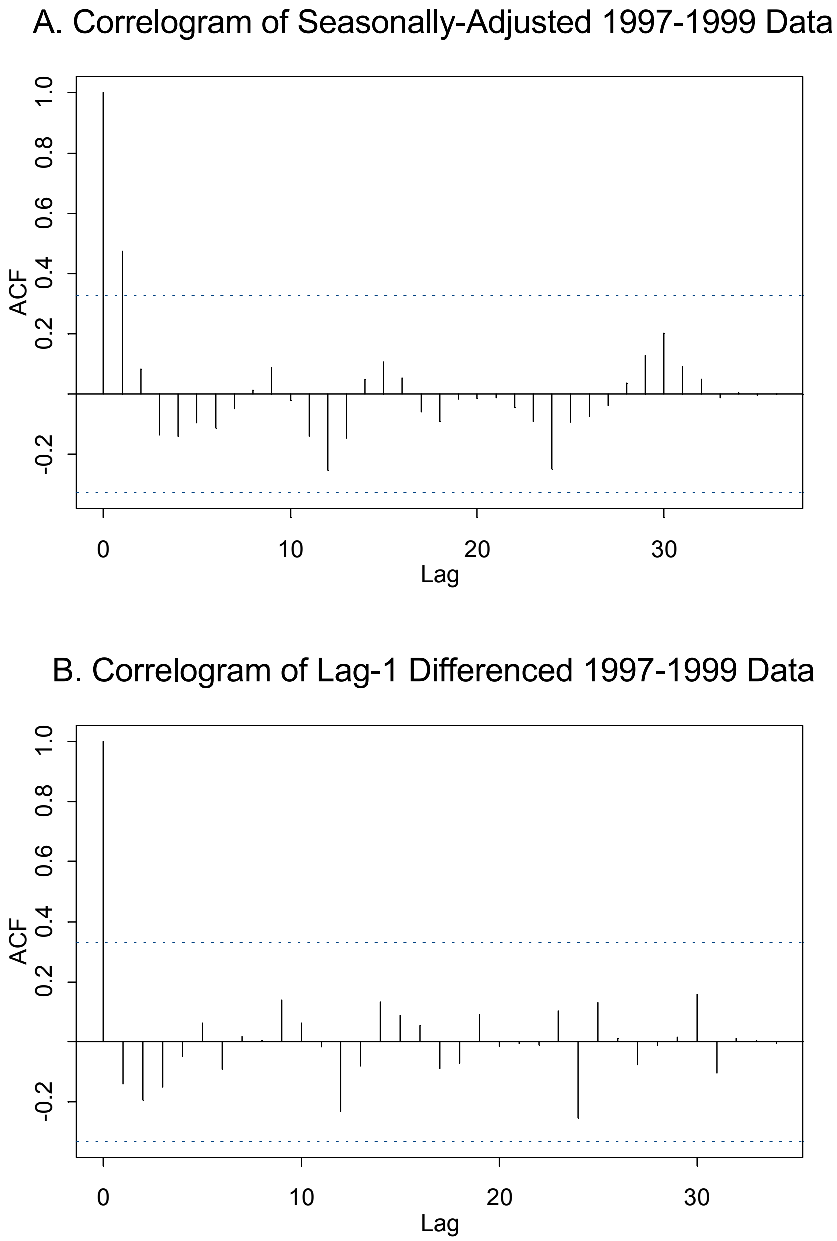
1. Compute the 1997, 1998 and 1997 annual means of the log-turbidity data (4.090, 4.407, 3.686)
2. Subtract from each observed log-turbidity value, its corresponding annual mean
3. Compute the twelve monthly means of the year-adjusted values computed in Step 3 (1.138, 0.0964, 1.281, 0.650, -0.304, -0.311, -0.8101, -1.111, -0.572, -0.250, -0.597, -0.167)
4. Compute the residual for each of the year-adjusted values and the corresponding monthly deviations computed in Step 3 by subtracting the appropriate annual and monthly mean from observed log-turbidity value. The following illustrates the computation of the residual for Month 1 of 1997 is computed as:

   Jan. 1997 log-turbidity value – 1999 mean – deviation from Jan. mean = Residual
   5.298 – 4.090 – 1.138 = 0.070

The next step is to examine the autocorrelation in the residuals by plotting their correlogram. This correlogram is shown in Fig. 2A. Although there still appears to be some seasonality in the data, it is now much weaker and, more importantly, only one significant autocorrelation remains. This correlation occurs between adjacent months (i.e., at lag1) and can be removed by subtracting from the residual of each month, the residual of the previous month. The resulting differences are called lag1-transformations or first-order differences. Since it is not possible to compute a lag1 difference for the first month (i.e., January 1997), the lag1 transformation always results in the loss of the first data point. Figure 2B shows the correlogram of the lag1-transformed residuals and confirms that the transformation has removed the remaining autocorrelation from the data.

At this point, one has two choices. The first is to use an **ARIMA model** to test hypotheses about the time series. These models are extremely flexible and can be used to model seasonal time-series with autocorrelations at lags of 1 and beyond (Brockwell and Davis 1987; Diggle 1990). However, ARIMA models are mathematically complex and require the assistance of a skilled time series analyst. Fortunately, a much simpler approach is available for cases such as the present one in which the only significant autocorrelation occurs at lag1. This approach leads to the construction of an upper one-sided 95% confidence interval on the 3-year mean of the log-turbidity. The one-sided alternative hypothesis that the three-year median is less than 150 NTU can be evaluated by checking whether the upper bound on the back-transformed confidence

Fig. 2. Correlograms of seasonally-adjusted, monthly log-transformed turbidity data from the Mermantau River, January 1997- December 1999. (A) Data prior to lag1 differencing. (B) Data after lag1-differencing.



A. Correlogram of Seasonally-Adjusted 1997-1999 Data

B. Correlogram of Lag-1 Differenced 1997-1999 Data interval is greater than 150. The log-scale upper bound is computed from the following formula:

$$UCL = \bar{y} + \left(t_{1-a,df} \times \frac{\sqrt{s^2_{adj}}}{\sqrt{nm}} \times \sqrt{\frac{1+\hat{f}_1}{1-\hat{f}_1}} \right) \qquad (2)$$

where, $\bar{y}$ = the mean log-turbidity

$t_{1-a,df}$ = the t-statistic value associated with p=1-α and df degrees of freedom
n = the number of years of data
m = the number of months in a year
df = ((n×m)-m)/3
s^2_{adj} = the variance of the seasonally adjusted log-turbidity
$\hat{f}_1$ = the estimated autocorrelation in the seasonally adjusted log-turbidity at lag1

The eleven steps required to obtain this confidence interval will now be illustrated with the data from Table 1.

1. Compute the square of each seasonally adjusted residual (i.e., square each value in column 5). The resulting squared values are given in Column 6.
2. Sum the squared residuals. The result is 8.348.
3. Compute the variance of the seasonally adjusted residuals:

   $$s^2_{adj} = \frac{1}{(nm)-m} \times sum\ of\ squared\ residuals$$

   = (1/((3×12)-12)) × 8.348
   = 0.348

   and,

   s_{adj} = $\sqrt{0.348}$ = 0.590
4. Multiply each residual by the value of the residual in the preceding month. These values are listed in column 7.
5. Sum the products from Step 4. The result is 3.955
6. Compute the autocorrelation between each log-turbidity and the log-turbidity in the preceding month by dividing the sum of the products of the lagged residuals by the sum of the squared seasonally adjusted residuals:

   $\hat{f}_1$ = 3.955/8.348 = 0.473

   note: this is the value of the lag1 autocorrelation shown in the correlogram in Fig. 2B
7. Compute the mean of the 36 monthly log-turbidity values (148.4/36). This value is 4.12.
8. Compute the degrees of freedom for the t-statistic = ((n×m)-m)/3 = ((3×12)-12)/3 = 8
9. Look up the value of the t-statistic associated with p=0.95 and df=8. The result is t=1.86.
10. Carry out the arithmetic in Eq. 2, substituting the values computed in steps 1-9,

    $$\text{UCL} = 4.12 + \left[1.86 \times \left(\frac{0.590}{6} \right) \times \sqrt{\frac{1+0.473}{1-0.473}} \right]$$

    = 4.12 + (1.86 × 0.098 × 1.672)
    = 4.427

11. Now back-transform the log-scale mean and the log-scale UCL to get the geometric mean and its upper 1-sided 95% confidence limit:
    Geometric Mean = Exp(4.12) = 61.68
    95% upper –sided confidence limit = exp(4.427) = 83.68

Having initially verified (Appendix C, Section C.2.5) that the 1980-2000 "population" of turbidity values fit a lognormal distribution, we can interpret the geometric mean and its 95% upper bound to be a valid estimate of the population upper 1-sided confidence interval on the population median. Since this interval does not include the criterion value, we can conclude that the sample evidence does not support the hypothesis that the mean/median turbidity in the Mermentau River was greater than or equal to 150 NTU during 1997-1999. Had we not log-transformed the data, our estimate of the mean would not have provided an unbiased estimate of the population median. Likewise, if we had not adjusted for autocorrelation, our estimate of the standard error of the geometric mean and its confidence interval also would have been biased. The procedure just described corrects for both problems without loss of data.

D.3 Nonparametric one-sample tests on means

Tests like the t-tests that require the sample data to have some specific parametric distributional form (e.g., the normal) are called **parametric tests.** When the data do not come from such a distribution, or when the data set is too small to tell what distribution it came from, or when a suitable normalizing transformation cannot be found, a class of tests called **nonparametric tests.** may be employed to compare the mean or median of a continuous water quality variable (e.g., turbidity) against a criterion value (e.g., the maximum allowable turbidity value in a stream). In this section we describe two such tests, the Wilcoxon signed ranks tests and Chen's modified t-test. These and other nonparametric tests useful for analyzing WQS data are described in Hollander and Wolfe 1999 and Millard and Neerchal 2001.

The formula for the exact Wilcoxon test statistic is described in Box 7. Like the t-test on log(X), the Wilcoxon signed-ranks test evaluates the null hypothesis that the median of the differences between the standard (M_0) and the sample values is zero. The sample value of W can be compared against the exact distribution of the Wilcoxon sign-ranks test statistic to determine the probability of obtaining a larger value under the null hypothesis (Hollander and Wolfe 1999). In practice, one usually uses statistical software (e.g., SAS; SPLUS EnvironmentalStats) to obtain the exact p-values.

Wilcoxon signed rank test depends on the following assumptions:

1. The variable must be measured on an interval scale; i.e. it can be either a count or a continuous random variable.
2. The sampling units from which the values of the variable were measured must be spatially and temporally independent
3. The frequency distribution of the variable values must be symmetric

Assumption 1 is verified simply by examining the measurement scale, assumption 2 is verified using the graphical methods illustrated in Sections C.2.4 and C.2.5, and assumption 3 is verified by constructing a frequency histogram of the sample data (e.g., Fig. 5a of Appendix C). When all the assumptions hold, the Wilcoxon signed ranks test is valid.

Box 7: Exact Wilcoxon Signed-Ranks Test for the Median

Assumptions

3. The population distribution of the attribute X is symmetric around the population median M.
4. The sample $X_1, X_2, \ldots, X_n$ is a random sample of *n* independent data points from the target population.

Let M_0 be the fixed criterion against which the population median M is compared. Consider each of the hypothesis pairs:

Case 1: $H_0 : M = M_0$ vs. $H_a : M > M_0$
Case 2: $H_0 : M = M_0$ vs. $H_a : M < M_0$

These are the steps for the exact Wilcoxon signed-ranks test.

Step 1. Select a significance level *a.*
Step 2. Calculate the difference $d_i = X_i - M_0$ for each of the *n* data points.
Step 3. Rank [d_i], the absolute value of the differences, from lowest to highest. That is, the smallest [d_i] will have rank 1, the second smallest is rank 2, …, the largest [d_i] is rank *n*. If there are ties, assign the average of the ranks which would otherwise have been assigned to the tied observations. For instance, if 3 observations are tied for Rank 2 then assign the average of ranks 2, 3 and 4. Thus, these 3 tied observations each will have rank (2+3+4)÷3=3. The next rank after these 3 observations will be rank 5.

Step 4. Determine *sign(d$_i$)*, where $sign(d_i) = \begin{cases} +1 & \text{if } d_i \geq 0 \ (i.e.,\ x_i \geq M_0) \\ -1 & \text{if } d_i < 0 \ (i.e.,\ x_i < M_0) \end{cases}$.

Step 5. Calculate the sum W of the ranks with a positive sign.

$$W = \sum_{i=1}^{n} sign(d_i) \times Rank(| d_i |)$$

Step 6. Use the table of critical values of the Signed Ranks Statistic to find the critical value W_a. Compare *W* with W_a.
Case 1: If $W < W_a$ then reject H_0. Otherwise, accept H_0.
Case 2: If $W > \frac{n(n+1)}{2} - W_a$ then reject H_0. Otherwise, accept H_0.

Although the symmetry assumption of the Wilcoxon signed ranks is not as restrictive as the normality assumption, it precludes the analysis of skewed (e.g., lognormal) distributions, a serious drawback for analysis of environmental data. Chen's modified t-test offers a useful alternative for the analysis of skewed data. When the data are either right-skewed or left-skewed, the traditional t-test does not have good power (i.e., the Type II error rate is inflated).

Chen (1995) developed a modified form of the t-test that requires an estimate of the skew from which an adjustment is made to the value of the t-statistic to account for the skew. Chen's test requires the same assumptions as the traditional t-test (see Appendix C, Section C.3.1) with the notable exception that normality is *not* assumed. Computational details of Chen's modified t-test are summarized in Box 8a. Box 8b provides a detailed example the application of the test to evaluate the one-sided alternative hypothesis that the mean of a right-skewed distribution of herbicide concentrations is greater than the criterion value.

D.4 One-sample tests on binomial proportions

The ambient water quality criteria for several "conventional" parameters (e.g., dissolved oxygen, pH, temperature) are written in terms of the percentage of exceedant sampling units in a sample. For example, if the acceptable range of pH values for a body of water is set at 6.5-9.0, then any aliquot of water with a pH outside of this range will fail the standard. Consider a sample in which 17 of 100 aliquots collected from a lake have ph values outside this range; the sample exceedance rate is 17%. If the criterion for pH specifies that no more than 10% of the sampling units fail the standard, this sample would appear to exceed the criterion. However, a statistically valid assessment of the sample estimate requires that some accounting of the uncertainty in the sample be incorporated into the estimate. The statistical procedures described in this section and the next provide a method for doing so.

When the object of a study is to determine if the proportion of sampling units exceeding some water quality standard is greater than the proportion of exceedances permitted by a regulatory criterion (e.g., 10%) and the product of the sample size and the sample proportion ($n \times p$) and the product $n \times (1-p)$ are both ≥ 5.0, the proportions z-test may be used (Box 9a). In order to use a proportions test to evaluate a water quality criterion, the measured continuous values (e.g., pollutant concentrations) of the individual sample units must be converted to dichotomous (i.e., acceptable vs. exceedant) scores. Typically, this is done by creating a new variable, say Y, which is scored as 1 for each exceedant concentration or 0 for each acceptable concentration. The proportion of exceedances in the sample is then computed as the mean of Y. The one-sample binomial proportions test may be applied to evaluate two-sided alternatives (H_a population proportion $\neq$ criterion proportion) or lower (H_a population proportion < criterion proportion) or upper (H_a population proportion > criterion proportion) one-sided alternative hypotheses.

When the number of sampling units in the sample is large [i.e., $n \times p > 5.0$ and $n \times (1-p) \geq 5.0$] and the sampling units have been independently sampled from the target population, the test statistic in Box 9a will be normally distributed with mean $n \times p$ and variance $n \times p \times q$. For $\alpha=0.05$, a value of $z \geq 1.645$ will result in rejection of the null hypothesis that the population proportion

Box 8-a: Chen's Modified One-Sample t-Test for the Mean
(One-sided Case)

Assumptions

5. X is a continuous random variable with population mean *μ* and variance s^2.
6. X_1, X_2, …, X_n is an independent sample of *n* individuals from the target population.

Let μ_0 be the fixed criterion against which the population mean *μ* is compared. Consider each of the 1-sided hypothesis pairs:

Case 1: $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$
Case 2: $H_0 : \mu = \mu_0$ vs. $H_a : \mu < \mu_0$

For either case, when the two assumptions hold, Chen's modified t-test for the null hypothesis vs. the alternative hypothesis can be performed as follows:

Step 1. Select the significance level *a*. (Typical values of *a* are 0.05 and 0.01.)
Step 2. Calculate the sample size, n, mean $\bar{x}$ and the sample variance s^2 (see Appendix C, Box 1-a).
Step 3. Compute the sample skewness:

$$\sqrt{\hat{b}_1} = \frac{\hat{m}_3}{\hat{s}^3}$$

Where: $\hat{m}_3 = \dfrac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (x_i - \bar{x})^3$

$$\hat{s}^3 = \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{3/2}$$

Step 4. Calculate the standard t-statistic.

$$t_0 = \frac{\bar{x} - m_0}{\sqrt{\dfrac{s^2}{n}}}$$

where $\bar{x}$ = the sample mean of the measured attribute, X
s^2 = the sample variance of the measured attribute, X
n = the number of sampling units in the sample
μ_0 = the fixed criterion against which the population mean is compared.

Box 8-a: Chen's Modified One-Sample t-Test for the Mean (Continued)

Step 5. Compute Chen's skew-adjusted t-statistic:

$$t_c = t_0 + a\left(1+2t_0^2\right)+4a^2\left(t_0+2t_0^3\right)$$

Where: $a = \dfrac{\sqrt{\hat{b}_1}}{6\sqrt{n}}$

Step 6. Use the table of percentage points of student's t-distribution to find the value $t_{1-a,\, n-1}$ such that (1-*a*)x100% of the Student t distribution with n-1 degrees of freedom is less than $t_{1-a,\, n-1}$.

Step 7. Compare t_c with $t_{1-a,\, n-1}$.
Case 1: If $t_c > t_{1-a,\, n-1}$ then reject H_0.
If $t_c = t_{1-a,\, n-1}$ then reject H_0.
Case 2: If $t_c < t_{1-a,\, n-1}$ then reject H_0.
If $t_c = t_{1-a,\, n-1}$ then reject H_0.

Box 8-b: Example for Performing Chen's One-Sample t-Test for the Mean

Consider a random sample of 15 fish taken from a river. One measurement of Chlorophenoxy herbicide concentration in liver tissue was obtained from each fish. The sample values of the tissue concentrations and some intermediate calculations required for the sample mean, variance, and skewness estimates are:

		SUM OF	SUM OF
HERBICIDE		SQUARED	CUBED
CONC.	DEVIATION	DEVIATIONS	DEVIATIONS
10	-165.5	27,390	-4,533,086
13	-162.5	53,797	-8,824,102
20	-155.5	77,977	-12,584,131
36	-139.5	97,437	-15,298,836
41	-134.5	115,527	-17,731,974
59	-116.5	129,100	-19,313,142
67	-108.5	140,872	-20,590,431
110	-65.5	145,162	-20,871,442
110	-65.5	149,452	-21,152,453
136	-39.5	151,013	-21,214,083
140	-35.5	152,273	-21,258,822
160	-15.5	152,513	-21,262,546
200	24.5	153,113	-21,247,840
230	54.5	156,084	-21,085,961
1300	1124.5	1,420,584	1,400,844,570
2632			

The sample mean (175.5) and variance (101,470.29) are computed by dividing the sums of the first column and the last entry in the third columns by the sample size, n (15) and n-1, respectively. The values in column 2 are computed by subtracting the mean fro each value in column 1. Using the equations in Box 8-a, the skew can be computed in three steps:

Step 1. $\hat{m}_3 = \left[\frac{15}{(15-1)(15-2)} \; x \; 1{,}400{,}844{,}570 \right]$

$= 115{,}454{,}22.8$

Step 2. $\hat{s}^{\,3} = (101{,}470.27)^{3/2}$

$= 32{,}322{,}752$

Step 3. $\sqrt{\hat{b}_1} = \frac{115{,}454{,}223}{32{,}322{,}743} = 3.572$

Box 8-b: Example for Performing Chen's One-Sample t-Test for the Mean (Continued)

Step 4. Compute the value of a (see Box 8-a):

$$a = \frac{3.572}{6\sqrt{15}} = 0.15373$$

It is desired to test the null hypothesis that the mean tissue concentration of the herbicide is no more than 100 µg/kg vs. the alternative the mean is greater than 100 µg/kg/

$$H_0 : m \le 100 \quad \text{vs.} \quad H_a : m > 100$$

Step 5. Thus the value of the usual t-statistic is computer as:

$$t_0 = \frac{175.5 - 100}{\sqrt{101{,}470.27 \big/ 15}} = 0.918$$

Step 6. Compute Chen's modified t-statistic:

$$\begin{aligned} t_c &= 0.918 + 0.15373(1 + 0.918^2) + \left[4 \times 0.15373^2 (0.918 + 2x0.918^3)\right] \\ &= 1.563 \end{aligned}$$

Note: the difference between the values of t_c and t_0 reflect the adjustment for the skew in the sample data.

Step 7. Set the desired a-level, e.g., 0.05; thus for the 1-sided case and n=15 we need to find the critical value of associated with df=14 and 1-a=0.95 in the table of percentage points of student's t-distribution. This value is 1.761.

Step 8. Comparing t_c = 1.563 to $t_{0(14,095)}$ = 1.761, we conclude that since the value of t_c is < t_0, the herbicide concentration in the sample of fish tissue support the null hypothesis that fish taken from the river attain the regulatory standard for chlorophenoxy herbicide.

Box 9-a: One-Sample Test for Proportions (Large Samples)

Assumptions

7. X is a dichotomous random variable taking on values 0, 1 denoting respectively the absence or presence of some attribute or condition (e.g., exceedance of some standard) observable for each sampling unit in the target population.
8. The sample X_1, X_2, …, X_n is a random sample of n sampling units from a population with a proportion, p , of its individuals with X=1
9. The values of n and p are such that both n×p and n× (1-p) = 5.0.

Let *p* denote the proportion of sampling units that exceed some criterion value. Let p_0 be the criterion value against which the population proportion p is compared. Consider each of the hypothesis pairs:

Case 1: $H_0 : p = p_0$ vs. $H_a : p > p_0$
Case 2: $H_0 : p = p_0$ vs. $H_a : p < p_0$

These are the steps for a large-sample (n>50) test for the population proportion.

Step 1. Select the significance level *a*. (Typically, *a* is 0.05 or 0.01.)
Step 2. Calculate the sample proportion *p* (see Box 1-a).
Step 3. Use the table of cumulative probabilities of the standard normal distribution to find the value $z_{1-a/2}$ such that (1-a)x100% of the standard normal distribution is below z_{1-a}.
Step 4. Calculate the test statistic z_c.

$$z_c = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where z_c = the standard normal test statistic
p = the proportion of exceedant sampling units in the sample
p_0 = the regulatory limit on the proportion of exceedances in the population.

Step 5. Compare z_c with z_{1-a}, the value of Z in the table of cumulative probabilities of the standard normal distribution that is associated with a probability = 1-*a*.
Case 1: If $z_c > z_{1-a}$ then reject H_0. Otherwise, accept H_0.
Case 2: If $z_c < z_{1-a}$ then reject H_0. Otherwise, accept H_0.

Box 9-b: Example for Performing a One-Sample Test for Proportions (Large Samples)

Consider a random sample of 100 monthly turbidity measurements from the Louisiana River between June 1980 and April 2000. The sample proportion of measurements that exceed 150 NTU is *p=0.19*. It is desired to test whether this sample proportion is indicative that the population proportion is larger than 0.15.

$$H_0 : p = 0.15 \quad vs. \quad H_a : p > 0.15$$

Step 1. The desired significance level is a=0.05.
Step 2. The sample proportion has been calculated to be *p=0.19*.
Step 3. The table of cumulative probabilities of the standard normal distribution yields $z_{0.95}$=1.645.
Step 4. Calculate the test statistic.

$$z = \frac{0.19 - 0.15}{\sqrt{\frac{(0.15)(0.85)}{100}}} = \frac{0.04}{\sqrt{0.001275}} = 1.12$$

Step 5. Since 1.2 < 1.645, then the null hypothesis cannot be rejected. Accept the null hypothesis and conclude that the true proportion of monthly turbidity measurements that exceed 150 NTU is less than or equal to 0.15.

of exceedances π is $\leq p_0$. Tables of the cumulative distribution of z are available in most elementary statistics texts and in all the standard commercial statistical software packages.

When the sample size criteria for the large sample normal approximation (Z-test) are not met, the behavior of the binomial becomes more discrete and is not well approximated by a continuous distribution such as the standard normal. For small samples, the exact binomial distribution must be used as the basis for both hypothesis tests and confidence intervals. The exact binomial test-statistic is simply the observed number of exceedances (r) out of the n sampling units in the sample. The p-value associated with r exceedances out of n sampling units is the upper-tailed cumulative binomial probability of observing r or more exceedances and can be computed with the formula in Box 10. The validity of the p-value for the exact binomial test rests on two assumptions: (1) the sampling units were selected from the target population by an independent, random process and (2) the underlying population exceedance probability is constant for every sampling unit in the population. Tables of the cumulative binomial P are available in many statistics texts (e.g., Hollander and Wolfe 1999) and from commercial statistical software packages (e.g., SAS; SPLUS EnvironmentalStats; StatExact).

The one-sided test of the null hypothesis that the observed proportion of exceedant sampling units comes from a population whose true proportion of exceedances is $\leq p_0$ is carried out by first finding the cumulative probability of observing ≥ r exceedances out of n sampling units (the p-value obtained from the equation in Box 10a). This p-value is then compared to the value of α specified by the analyst; if $p < \alpha$, she rejects H_0, otherwise she accepts H_0. Because of the discreteness of the binomial distribution, the actual α will generally be somewhat lower than the α specified in the DQOs. For example, if an investigator specifies $p_0 = 0.30$ and has a sample of n=17 with 11 exceedant sampling units and wants to test against the one-sided upper-tail alternative with $\alpha=0.05$, he will actually need to use $\alpha=0.0403$. This is because given n=17, the closest that one can actually come to p=0.05 (without exceeding it) is the cumulative probability of observing ≥ nine exceedances in the sample; i.e., p= 0.0403. This of course tells us that any number of exceedances ≥ 9 will cause rejection of H_0 for a sample size of n=17. The probability of observing ≥ 11 exceedances in a sample of 17 sampling units from a population with $p_0 = 0.30$ is 0.0032, which is $< \alpha=0.0403$; thus H_0 is rejected in this case.

D.5 Controlling Type I and II error rates for exact binomial tests

Although all of the interrelationships among n, power, and α described earlier hold for exact binomial tests, the relationships are complicated by the discreteness of the distribution. Table 2 summarizes these relationships for samples containing from 1 to 10 exceedances when the criterion level (p_0) is 10% exceedances, the sample size is fixed (n=10) and five different α-levels are specified. The following are the exact binomial probabilities of observing at least r exceedances out of a sample of 10 sampling units from a target population in which the true proportion of exceedances is 0.10:

Box 10-a: Exact Binomial Test for Proportions of Exceedances (Listing Case)

Binomial Distribution

Suppose there are *n* independent observations of a trait that has only two possible values (say *success* or *failure*). Let the probability p of observing a *success* be the same for all observations. Then the total number of successes, X, among the n observations has a binomial probability distribution, where the probability of observing *k* successes in a sample of size *n* is given by:

$$\Pr(X = k) = \left(\frac{n!}{k!(n-k)!} \right) p^{k}(1-p)^{n-k} \quad \text{where } k = 0, 1, 2, 3, ..., n$$

To obtain the probability of observing a range of values for X, add up the probabilities of observing each of the values in the range. Tables of the cumulative probabilities are available in most statistics texts (e.g., Hollander and Wolfe, 1999) and from commercial statistics software packages (e.g., SAS, SPLUS EnvironmentalStats, StatExact).

Assumptions:

1. The sample is selected from the target population by an independent, random process.
2. The attribute of interest is exceedance of a specified criterion.
3. The underlying population probability of exceedance, p, is constant for every unit in the population.

Then the total number of exceedances in a random sample of size *n* is a binomial random variable.

In order to list a body of water as being impaired it ;has to be demonstrated that the population proportion of exceedances is greater than the regulatory limit. This leads to the null hypothesis that the true proportion of exceedances in the population is less than or equal to a standard maximum allowable proportion of exceedances, p_0. Following are the steps for performing the exact binomial test for the listing case.

Step 1. Specify the desired significance level a (usually 0.05 or 0.01).

Step 2. The exact binomial test statistic is the observed number of exceedances, *r*, in the sample.

Step 3. Calculate the p-value, *P*, the upper-tailed cumulative probability of observing at least *r* exceedances among *n* sampling units, when the true proportion is assumed to be p_0.

$$\Pr(X \geq r) = P = \sum_{k=r}^{n} \left(\frac{n!}{k!(n-k)!} \right) p_0^{k}(1-p_0)^{n-k}$$

where P = the upper-tailed cumulative binomial probability
p_0 = the regulatory limit on the proportion of exceedances in the population
n = the sample size
r = the observed number of exceedant sampling units in the sample.

Table of the cumulative probabilities are available in most statistics tests (e.g., Hollander and Wolfe, 1999) and from commercial statistics software packages (e.g., SAS, SPLUS, EnvironmentalStats, StatExact).

Step 4. Compare the p-value, *P*, with the desired significance level a.
If *P*<a then reject the null hypothesis; otherwise, accept the null hypothesis.

Box 10-b: Exact Binomial Test for Proportions of Exceedances (Listing Case)

Consider a random sample of 10 monthly turbidity measurements from the Mermentau River between June 1980 and April 2000. The measurements (in NTU) are: 34, 58, 87, 145, 14, 38, 62, 95, 160, 320. Suppose that the maximum allowable proportion of exceedances is 0.15. Based on the sample estimate (i.e., 2/10 -= 0.20) the proportion of exceedances, is the Mermentau River impaired with respect to the criterion that the proportion should be =0.15?

$$H_0 : \text{p} = 0.15 \quad \textit{vs.} \quad H_a : \text{p} > 0.15$$

Step 1. Let the desired a=0.05.
Step 2. There are *r=2* exceedances in the sample.
Step 3. Calculate the p-value.

$$\begin{aligned}
\Pr(X \geq 2) = P &= \sum_{k=2}^{10} \left(\frac{10!}{k!(10-k)!} \right) (0.15)^{k} (0.85)^{10-k} \\
&= \left(\frac{10!}{2!(10-2)!} \right) (0.15)^{2} (0.85)^{10-2} + \left(\frac{10!}{3!(10-3)!} \right) (0.15)^{3} (0.85)^{10-3} \\
&\quad + \ldots + \left(\frac{10!}{10!(10-10)!} \right) (0.15)^{10} (0.85)^{10-10} \\
&= 0.2759 + 0.1298 + \ldots + (0.15)^{10} \\
&= 0.4557\,.
\end{aligned}$$

Step 4. Since 0.4557 > 0.05, then the null hypothesis cannot be rejected. Accept the null hypothesis and conclude that the sample does not provide sufficient evidence that the Mermentau River is impaired.

Table 2. Exact Binomial probabilities, and Type II error rates for n=10 and 5 different prespecified Type I error rates for listing water bodies.

SAMPLE SIZE	NUMBER OF EXCEEDANCES	SPECIFIED TYPE I	ACTUAL TYPE 1	TYPE II ERROR	LOWER 1-SIDED EXACT 95% CI	MIN. NO. TO REJECT
10	1	0.05	0.0128	0.9872	0.100 (0.005, 1.000)	4
		0.20	0.0702	0.9298	0.100 (0.022, 1.000)	3
		0.25	0.0702	0.9298	0.100 (0.028, 1.000)	3
		0.30	0.2639	0.7361	0.100 (0.035, 1.000)	2
		0.35	0.2639	0.7361	0.100 (0.042, 1.000)	2
10	2	0.05	0.0128	0.8791	0.200 (0.037, 1.000)	4
		0.20	0.0702	0.6778	0.200 (0.083, 1.000)	3
		0.25	0.0702	0.6778	0.200 (0.096, 1.000)	3
		0.30	0.2639	0.3758	0.200 (0.109, 1.000)*	2
		0.35	0.2639	0.3758	0.200 (0.122, 1.000)*	2
10	3	0.05	0.0128	0.6496	0.300 (0.087, 1.000)	4
		0.20	0.0702	0.3828	0.300 (0.158, 1.000)*	3
		0.25	0.0702	0.3828	0.300 (0.176, 1.000)*	3
		0.30	0.2639	0.1493	0.300 (0.193, 1.000)*	2
		0.35	0.2639	0.1493	0.300 (0.209, 1.000)*	2
10	4	0.05	0.0128	0.3823	0.400 (0.150, 1.000)*	4
		0.20	0.0702	0.1673	0.400 (0.239, 1.000)*	3
		0.25	0.0702	0.1673	0.400 (0.261, 1.000)*	3
		0.30	0.2639	0.0464	0.400 (0.281, 1.000)*	2
		0.35	0.2639	0.0464	0.400 (0.300, 1.000)*	2
10	5	0.05	0.0128	0.1719	0.500 (0.222, 1.000)*	4
		0.20	0.0702	0.0547	0.500 (0.327, 1.000)*	3
		0.25	0.0702	0.0547	0.500 (0.351, 1.000)*	3
		0.30	0.2639	0.0107	0.500 (0.373, 1.000)*	2
		0.35	0.2639	0.0107	0.500 (0.393, 1.000)*	2
10	6	0.05	0.0128	0.0548	0.600 (0.304, 1.000)*	4
		0.20	0.0702	0.0123	0.600 (0.419, 1.000)*	3
		0.25	0.0702	0.0123	0.600 (0.445, 1.000)*	3
		0.30	0.2639	0.0017	0.600 (0.468, 1.000)*	2
		0.35	0.2639	0.0017	0.600 (0.489, 1.000)*	2
10	7	0.05	0.0128	0.0106	0.700 (0.393, 1.000)*	4
		0.20	0.0702	0.0016	0.700 (0.516, 1.000)*	3
		0.25	0.0702	0.0016	0.700 (0.542, 1.000)*	3
		0.30	0.2639	0.0001	0.700 (0.566, 1.000)*	2
		0.35	0.2639	0.0001	0.700 (0.587, 1.000)*	2

* PROPORTION OF EXCEEDANCES SIGNIFICANTLY > 0.10

Table 2. (continued)

SAMPLE SIZE	NUMBER OF EXCEEDANCES	SPECIFIED TYPE I	ACTUAL TYPE 1	TYPE II ERROR	LOWER 1-SIDED EXACT 95% CI	MIN. NO. TO REJECT
10	8	0.05	0.0128	0.0009	0.800 (0.493, 1.000)*	4
		0.20	0.0702	0.0001	0.800 (0.619, 1.000)*	3
		0.25	0.0702	0.0001	0.800 (0.645, 1.000)*	3
		0.30	0.2639	0.0000	0.800 (0.667, 1.000)*	2
		0.35	0.2639	0.0000	0.800 (0.687, 1.000)*	2
10	9	0.05	0.0128	0.0000	0.900 (0.606, 1.000)*	4
		0.20	0.0702	0.0000	0.900 (0.729, 1.000)*	3
		0.25	0.0702	0.0000	0.900 (0.753, 1.000)*	3
		0.30	0.2639	0.0000	0.900 (0.773, 1.000)*	2
		0.35	0.2639	0.0000	0.900 (0.791, 1.000)*	2
10	10	0.05	0.0128	0.0000	1.000 (0.741, 1.000)*	4
		0.20	0.0702	0.0000	1.000 (0.851, 1.000)*	3
		0.25	0.0702	0.0000	1.000 (0.871, 1.000)*	3
		0.30	0.2639	0.0000	1.000 (0.887, 1.000)*	2
		0.35	0.2639	0.0000	1.000 (0.900, 1.000)*	2

* PROPORTION OF EXCEEDANCES SIGNIFICANTLY > 0.10

r	Pr(no. exceedances ≥ r)
1	0.6513
2	0.2639
3	0.0702
4	0.0128
5	0.0016

Thus if an investigator wants to specify $\alpha=0.05$, he has to consider an actual α that is slightly larger (0.0702) or slightly smaller (0.0128). This is because there are only 10 possible exceedance values >0 for a sample of size 10. The conservative choice will always be to choose the probability closest to, but less than, the desired α. The desired α-levels are specified in the third column of Table 2 and the corresponding actual α-levels are in column four. These values are fixed for all samples of n=10, regardless of how many exceedances (column two) they might contain. Note that since at least 4 exceedances are required to achieve a probability ≤ an actual α =0.0128, only samples with ≥ four exceedances will result in rejection of H_0. Similarly, at least three exceedances will be required for rejection at specified α-levels of 0.20 and 0.25, while a minimum of two exceedances will be required to reject at α-levels of 0.30 and 0.35. Because the numbers of exceedances that are required to reject the null hypothesis in a sample of ten sampling units (column seven) are specific to the α-level, they do not change with the number of exceedances actually observed in the sample.

The observed proportion of exceedances in a sample (r/n) has a strong effect on the Type II error probability. The difference between the observed proportion and the criterion value ($p_0=0.10$) is actually an effect size measure. Recall that for fixed sample size, α-level, and variance, the effect size will determine the observed power and hence the observed Type II error probability. The variance of a binomial response is [p ×(1-p)], thus the variance changes with the number of exceedances in the sample. However, for proportions in the range of 0.20 to 0.80, the differences in the variance are quite small. Therefore most of the change in Type II error for a given α-level in Table 2 is due to the increasing effect size. For example, the Type II error rate for the specified $\alpha=0.05$ in a sample with 3 exceedances is 0.6496, but in a sample with 4 exceedances it is only 0.3823; the corresponding sample variances are 0.21 and 0.24. Note in Table 2 that the Type II error rates (for all α-levels) drop precipitously between r=1 and r=5 exceedances. This corresponds to the gray region in Figure 16 and indicates that a sample of size of 10 has reasonably good power when the effect size is ≥ 0.40 (i.e., $\delta \geq 5/10$ – the criterion value). In other words, our telescope (sample with n=10 and $\alpha=0.0128$) does not have sufficient power for us to distinguish clearly (i.e., with Type II error < 0.20) two mountains (test statistic distributions) whose centers (proportions of exceedances) are any closer than 0.40.

The lower one-sided 100(1-α)% exact binomial confidence intervals (column 6) provide an estimate of the proportion of the exceedances in the target water body. Whenever the lower bound on this estimate is > 0.10, we reject H_0 and conclude that the sample evidence supports the alternative hypothesis that the water exceeds the acceptable standard. All such confidence intervals are marked with an asterisk. Notice that although the criterion is 0.10, when the specified α-level is 0.05, samples with 10%, 20% and 30% exceedances do not have lower bounds > 10%. This reflects the uncertainty in the estimate, but it also clearly demonstrates the high likelihood of erroneously accepting that such bodies of water meet the water quality standard when, in fact, they do not. The minimum number of exceedances required for a particular 100(1-α)% lower bound to be > than 10% is exactly the same as the minimum number required by the exact binomial test to reject H_0 for a specified α-level. For example, the first 95% confidence interval with an asterisk occurs when there are four exceedances in the sample. Similarly, the first 70% confidence interval with an asterisk occurs when there are only two exceedances. Thus regardless of which statistical tool one chooses (i.e., confidence intervals or exact binomial tests), the decision remains the same. Furthermore, both tools are subject to the same Type I and II error rate problems when small sample sizes are used.

The exact binomial test may also be employed for the complementary process of delisting a body of water that was previously found to be impaired (Box 11a). Using the same criterion level (i.e., the population exceedance rate must be < 10%), a table similar to Table 2 may be constructed to illustrate the delisting problem. Table 3 summarizes relationships among n, α, β for the delisting scenario for sample sizes of 22-28. Regardless of the specified α, it will never be possible to delist a previously listed water with a sample size less than 22. Recall that in the listing case,

$$H_0 : p \leq p_0 \quad vs. \quad H_a : p > p_0$$

Where p = observed proportion of exceedances

p_0 = standard maximum allowable proportion of exceedances

Thus, we reject the null hypothesis and list any water whose proportion of exceedances > 0.10 (p_0). A statistician would say that the rejection region for this decision scenario is between 0.10 and 1.0. By contrast, for the delisting case, the null and the alternative hypotheses for the listing scenario are essentially flip-flopped:

$$H_0 : p \geq p_0 \quad vs. \quad H_a : p < p_0$$

Therefore, we will only delist a listed a body of water if the proportion of exceedances in the sample is < 0.10. The statistical rejection region in this case is much smaller (0-0.10) than for the listing case (0.10-1.0). This requires a much more powerful telescope; i.e. much larger sample sizes. Consequently, it will always be much more difficult to delist a body of water than it was to list it in the first place.

The third column from the left in Table 3 is the compliment of the Type I error rate (i.e., 1-α); column three is the Type II error rate (β), and the last column is the maximum number of exceedances out of a sample of n sampling units which can be tolerated if the water is to be delisted. The fourth column is just the value in the last column divided by n and is therefore the maximum tolerable exceedance rate for a sample of given n and α. Comparing Tables 2 and 3, generally speaking, there is a much greater likelihood that a "bad", previously unlisted water, will not be listed than that a "bad" previously listed water will be delisted.

In the context of water quality attainment decisions, both Type I and Type II errors are cause for concern. For the landowner, an erroneous listing of the water (Type I error) on his property may result in severe financial hardship. On the other hand, the failure to list and subsequently restrict the use of a impaired water (Type II error) may have dire public health consequences. Therefore, it is difficult to argue that either type of error is more important than the other. This would seem to justify the specification of balanced α- and β-levels. Unfortunately, as we have demonstrated

Box 11-a: Exact Binomial Test for Proportions of Exceedances (Delisting Case)

Assumptions

10. The sample is selected from the target population by an independent, random process.
11. The attribute of interest is exceedance of a specified criterion.
12. The underlying population probability of exceedance, p, is constant for every unit in the population.

Then the total number of exceedances in a random sample size of *n* is a binomial random variable.

The exact binomial test may also be employed for the complementary process of delisting a body of water that has been previously found to e noncompliant. In this case, the null hypothesis that the true proportion of exceedances in the population is greater than or equal to a standard maximum allowable proportion of exceedances, p_0. The steps for performing the test are similar to the listing case.

Step 1. Determine the significance level *a* (usually 0.05 or 0.01.)
Step 2. The exact binomial test statistic is the observed number of exceedances, *r*, in the sample.
Step 3. Calculate the p-value, *P*, the *lower-tailed* cumulative probability of observing *r* or fewer exceedances among *n* sampling units, when the true proportion is assumed to be p_0.

$$\Pr(X \le r) = P = \sum_{k=0}^{r} \left(\frac{n!}{k!(n-k)!} \right) p_0^k (1-p_0)^{n-k}$$

where P = the lower-tailed cumulative binomial probability
p_0 = the standard maximum allowable proportion of exceedances in the population
n = the sample size
r = the observed number of exceedant sampling units in the sample.

Step 4. Compare the p-value, *P*, with the desired significance level *a*.
If $P<a$ then reject the null hypothesis. Otherwise, accept the alternative hypothesis.

TABLE 3. ERROR RATES FOR SAMPLES CONTAINING THE MAXIMUM NUMBER OF ALLOWABLE EXCEEDANCES TO DELIST FOR SAMPLE SIZES 22-28

	SPECIFIED	PROB. KEEPING	PROB. KEEPING		MAX. NO.
	TYPE I	BAD WATER ON	GOOD WATER ON	OBSERVED	EXCEEDANCES
N	ERROR RATE	LIST (1-α)*	ON LIST (β)	EXCEEDANCE	TO DELIST
22	0.05	0.9015	0.6406	0.0000	0
	0.20	0.9015	0.6406	0.0000	0
	0.25	0.9015	0.6406	0.0000	0
	0.30	0.9015	0.6406	0.0000	0
	0.35	0.6608	0.2641	0.0455	1
23	0.05	0.9114	0.6406	0.0000	0
	0.20	0.9114	0.6406	0.0000	0
	0.25	0.9114	0.6406	0.0000	0
	0.30	0.9114	0.6406	0.0000	0
	0.35	0.6849	0.2642	0.0435	1
24	0.05	0.9202	0.6399	0.0000	0
	0.20	0.9202	0.6399	0.0000	0
	0.25	0.9202	0.6399	0.0000	0
	0.30	0.7075	0.2642	0.0417	1
	0.35	0.7075	0.2642	0.0417	1
25	0.05	0.9282	0.6396	0.0000	0
	0.20	0.9282	0.6396	0.0000	0
	0.25	0.9282	0.6396	0.0000	0
	0.30	0.7288	0.2642	0.0400	1
	0.35	0.7288	0.2642	0.0400	1
26	0.05	0.9354	0.6393	0.0000	0
	0.20	0.9354	0.6393	0.0000	0
	0.25	0.9354	0.6393	0.0000	0
	0.30	0.7487	0.2642	0.0385	1
	0.35	0.7487	0.2642	0.0385	1
27	0.05	0.9419	0.6391	0.0000	0
	0.20	0.9419	0.6391	0.0000	0
	0.25	0.7674	0.2642	0.0370	1
	0.30	0.7674	0.2642	0.0370	1
	0.35	0.7674	0.2642	0.0370	1
28	0.05	0.9477	0.6388	0.0000	0
	0.20	0.9477	0.6388	0.0000	0
	0.25	0.7849	0.2642	0.0357	1
	0.30	0.7849	0.2642	0.0357	1
	0.35	0.7849	0.2642	0.0357	1

* FOR N <29 IT IS NOT POSSIBLE FOR ACTUAL α TO BE ≤ 0.05

previously, simultaneous control of α and β to levels ≈ 0.05 requires unrealistically large sample sizes.

It has become standard practice in the scientific literature to specify an α-level of 0.05; in those papers where it is considered, the maximum acceptable β-level is generally set at 0.20. Freedman et al. (1991) provide the following explanation of the origin of the 0.05 α-level:

> "R. A. Fisher was the first to use such tables [i.e., tables of test statistics associated with probabilities of 0.05 and 0.01] and it seems to have been his idea to lay them out this way. There is a limited amount of room on a page. Once the number of α-levels was limited to [two values], 5% and 1% stood out as nice round numbers and they soon acquired a magical life of their own. With computers everywhere, this type of table is obsolete. So are the 5% and 1% α-levels."

Smith et al. (2001) have suggested balancing the error rates at moderate levels (e.g., ≤ 0.15). This requires the investigator to specify both α and β levels, *a priori*, at step 6 of the DQO process. Because of the discrete nature of the binomial distribution, α and β levels can only be specified subject to the specification of a minimum number of exceedances required to reject H_0 (what Smith et al. call the "cut-off"). In the present example H_0: p =0.10. As explained previously, the Type II error rate cannot be set without first specifying a minimum effect size. Smith et al. propose that a population exceedance rate of 0.25, "indicates severe problems and represents the minimum violation [rate] we would almost always want to detect". Thus they recommend specifying p=0.25 for the population under $H_{a,}$ which is equivalent to specifying a minimum effect size of 0.15. Employing these specifications, a table of balanced Type I and Type II error rates for sample sizes less than 50 can be computed (Table 4).

Due to the discreteness of the binomial, the error rates can only be balanced for one sample size per each unique cut-off value (Table 4, column 2). The error rates for sample sizes less than 28 are probably too high to be acceptable to most regulators. As with so many investigations of sample size, power and precision, the results in Table 4 seem to suggest than n=30 is the "magic number". Clearly, the use of balanced error rates does not solve the problems associated with small sample sizes; however, it does remove the inequality associated with protection of one error rate at the expense of the other. Moreover, the balanced error presentation clearly reveals the risks associated with making decisions based on small sample sizes.

A similar table of balanced error rates can be computed for the delisting scenario (Table 5). Although the criterion for listing is 10% exceedance, we will adopt the reasoning of Smith et al. (2001) and specify that the exceedance rate should be $\geq$ 20% (i.e., H_0: P$\geq$ 20%) to justify keeping a body of water on the list; whereas, any body of water with <10% exceedance (i.e., H_a: P< 10%) will be taken off the list. As pointed out earlier, the mathematics involved with flipping the null and the alternative hypotheses result in a very small rejection region, so that it will be hard to reject the null hypothesis except when we have large samples. The practical consequence of all this is that much more evidence is required to delist than to list. The results in Table 5 demonstrate that if the minimally acceptable balanced error rates are to be held to $\leq$ 15%, the minimum sample size for delisting a body of water must be 59 (i.e., slightly more than double

Table 4. Balanced Type I and II error rates for the exact binomial for eight sample sizes for listing water bodies.

	MIN. NO.	TYPE I	TYPE II	
SAMPLE	TO REJECT	ERROR	ERROR	
SIZE	(CUTOFF)	PROB	PROB	POWER(%)
4	1	0.34	0.32	68.4
10	2	0.26	0.24	75.6
16	3	0.21	0.20	80.3
22	4	0.17	0.16	83.8
28	5	0.14	0.14	86.5
34	6	0.12	0.11	88.6
40	7	0.10	0.10	90.4
46	8	0.08	0.08	91.8

Table 5. Balanced Type I and II error rates for the exact binomial for thirteen sample sizes for delisting water bodies.

SAMPLE SIZE	MAX. NO. TO REJECT (CUTOFF)	TYPE I ERROR PROB	TYPE II ERROR PROB	POWER(%)
25	3	0.23	0.24	0.7636
32	4	0.20	0.21	0.7885
39	5	0.18	0.19	0.8097
46	6	0.16	0.17	0.8281
52	7	0.16	0.14	0.8560
59	8	0.14	0.13	0.8690
66	9	0.12	0.12	0.8800
73	10	0.11	0.11	0.8900
80	11	0.10	0.10	0.9000
87	12	0.09	0.09	0.9080
94	13	0.08	0.08	0.9150
100	14	0.08	0.07	0.9270
108	15	0.07	0.07	0.9290

the number to needed list with the same error rates). An example of the computation of exact binomial proportions for the delisting case is presented in Box 11b.

With respect to employing exact binomial procedures to support water quality attainment decisions, the preceding discussion indicates the following:

1. The sample of sampling units should be obtained from a design that insures independent and representative sampling of a target population clearly bounded in space and time.
2. Tests with balanced Type I and II errors are preferable to tests designed primarily to minimize Type I error rates.
3. Balanced Type I and Type II error rates should be less than 0.15.
4. The minimum effect size employed to compute balanced error rates should be based on careful consideration of the sampling costs and consequences of the Type II error.
5. In order to meet the requirements of 1 and 3, the sample should contain at least 28 sampling units for the listing case and 59 for the delisting case.

6. The binomial test and associated confidence intervals will be valid only if the exceedance rate is constant throughout the target population; i.e., there should be any spatial or temporal heterogeneity in the distribution of the exceedances.

The binomial model will lead to appropriate decisions only to the extent that the sample of n sampling units represents the true spatial and temporal variability of the body of water in question. The binomial model is no panacea for inadequate sample size, poor (i.e. nonrandom or restricted) sampling designs, or populations with extreme heterogeneity of exceedance rates. For heterogeneous populations, it will probably be more prudent to base water quality attainment decisions on tests comparing sample means or medians against the appropriate pollutant concentration or biological abundance criterion.

D.6. Estimation of the total exceedances in a 3-year period

Both acute and the chronic water quality criteria may be exceeded once per 3-year assessment period without consequence; however, waters with two or more exceedances of either standard will be listed as non-attainment waters. This is equivalent to requiring the total number of exceedances to be less than 2 for any given 3-year assessment period. Because 3-year attainment criteria are based on sampling from local bodies of water within the 3-year period, estimates of the total number of exceedances are subject to sampling error and other sources of uncertainty. Thus, it is necessary to quantify this error so that the appropriate confidence intervals and hypothesis tests can be constructed. A strategy for doing so is presented in this section.

We will take as an example 36 monthly sampling units collected from a particular river reach during 1997 –1999 for the assessment of the acute selenite criterion. The selenite concentration in each monthly sampling unit is compared against the selenite CMC (186 μg/l) and scored 1 if it exceeds the CMC; zero, otherwise. The sum of these scores is the sample estimate of the total number of exceedances in the population. But what is the population? Recall that the CMC is based on the hourly mean of repeated measurements taken within a 24-hour period. The total

Box 11-b: Exact Binomial Test for Proportions of Exceedances (Delisting Case)

Consider the example in Box 10-b. Suppose instead that the Mermentau River has been listed as noncompliant. Suppose also that the maximum allowable proportion of exceedances is 0.25. Based on the sample proportion of exceedance of 0.2, can the Mermentau River be deemed compliant and delisted?

$$H_0 : \text{p} = 0.25 \quad \textit{vs.} \quad H_0 : \text{p} < 0.25$$

Step 1. Let a=0.05.
Step 2. There are *r*=2 exceedances in the sample.
Step 3. Calculate the p-value.

$$\begin{aligned}
\Pr(X \le 2) = P &= \sum_{k=0}^{2}\left(\frac{10!}{k!(10-k)!}\right)(0.25)^k(0.75)^{10-k} \\
&= \left(\frac{10!}{0!(10-0)!}\right)(0.25)^0(0.75)^{10-0} + \left(\frac{10!}{1!(10-1)!}\right)(0.25)(0.75)^{10-1} \\
&\quad + \left(\frac{10!}{10!(10-2)!}\right)(0.25)^2(0.75)^{10-2} \\
&= 0.0563 + 0.1877 + 0.2816 \\
&= 0.5256\,.
\end{aligned}$$

Step 4. Since 0.5256 > 0.05 then the null hypothesis cannot be rejected. Accept the null hypothesis and conclude that the Mermentau River cannot be delisted.

number of such means that could be computed for a 3-year period is just the number of days in 3 years ≈ 3×365 = 1095 days. Thus, the total number of exceedances in the population must be between zero and 1095. Since we are actually categorizing the day upon which a particular monthly assessment was made as an exceedant or a compliant day, it is the individual days of the 3-year period that are the sampling units for the estimates of the total numbers of exceedant days.

Based on these definitions of the population (i.e., 1095 days) and the sample (36 days randomly sampled out of the target population), and the dichotomous outcome (the daily mean is classified as exceeding/not exceeding the CMC) it might at first appear that this is simply a problem in binomial proportions in which the sample proportion of exceedant days is compared to the criterion value of 2/1095 (i.e., proportion of CMC exceedances must be < 0.0018). However, there is one crucial difference; whereas, the target populations in the previous applications of the binomial distribution were essentially infinite (e.g., the number of 1-liter aliquots of water flowing through a river reach in a month), in this case, we have a finite target population. The hypergeometric distribution is the appropriate probability model for estimation of dichotomous proportions from a finite population (Cochran 1977; Thompson 1992). The probability of observing x number of exceedances in a sample of size n, given that it was randomly selected from a target population of fixed size N, containing exactly k exceedances is,

$$\Pr(X=x)=\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}= \tag{3}$$

$$\frac{\dfrac{k!}{x!(k-x)!}\,\dfrac{(N-k)!}{(n-x)!(N-k-n+x)!}}{\dfrac{N!}{n!(N-n)!}}$$

Where, x = the number of exceedant days in the sample of n days
n = the number of days sampled
N = the total number days in a 3-year assessment period = 1095
k = the number of exceedant days in the population

A related expression provides an estimate of the probability of observing ≤ x exceedant days in the sample. The expression is just the sum of all of the individual probabilities of observing 0 to x exceedances, each calculated with equation 3. This expression is called the cumulative distribution function (CDF),

$$\Pr(X \le x) = \sum_{r=0}^{x} \frac{\begin{pmatrix} k \\ r \end{pmatrix}\begin{pmatrix} N-k \\ n-r \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \tag{4}$$

The probability of observing more than x exceedances (i.e., = x+1) is just 1 minus CDF. This upper tailed hypergeometric probability provides the basis for deciding between the following two hypotheses:

H_0: the number of exceedances in the target population = k
H_a: the number of exceedances in the target population > k.

Or equivalently:

H_0: the proportion of exceedances in the target population = k/N
H_a: the proportion of exceedances in the target population > k/N.

For any given x, the resulting upper tailed p-value is compared to α and if it is smaller, the null hypothesis is rejected; otherwise it is accepted. Here, we want to choose x such that our specified a rate is not exceeded. This is the finite sampling counterpart to the exact binomial test for binary outcomes from an infinite population. Like the binomial, exact 100×(1-α)% confidence intervals can be formed for the total number of exceedances in the population or the corresponding proportion of exceedances. The lower 1-sided exact confidence limit on the number of exceedant days in the population is computed as the smallest k such that,

$$\Pr(X \ge x) = 1 - \sum_{r=0}^{x-1} \frac{\begin{pmatrix} k \\ r \end{pmatrix}\begin{pmatrix} N-k \\ n-r \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} > a \tag{5}$$

and the upper 1-sided exact confidence limit is computed as the largest k such that,

$$\Pr(X \le x) = \sum_{r=0}^{x} \frac{\begin{pmatrix} k \\ r \end{pmatrix}\begin{pmatrix} N-k \\ n-r \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} > a \tag{6}$$

The values of k that satisfy equations 5 and 6 are the lower and upper bounds on the total number of exceedant days during the 3-year monitoring period. To obtain the corresponding bounds on the proportion of exceedant days in the 3-year period, we simply divide each of the k values by N (i.e., 1095). Two-sided confidence intervals can be formed by replacing α with α/2 in equations 5 and 6 (Buonaccorsi 1987; Wendell and Schmee 2001).

As an example, consider the selenite monitoring program in which samples were collected monthly for 3 years. The CMC criteria require that there be no more than 1 CMC exceedance by any of the 1095 daily means during that period. What is the probability of obtaining 0, 1, or = 2 exceedant days in the sample of 36 days, if the true population had only 1 exceedant day out of 1095? We can compute the first two probabilities by substituting x=0 and x=1 into equation 3, specifying n=36, N=1095 and k=1; the third probability (which must be 0) is just one minus the sum of the first two (i.e., Equation 4):

$$\Pr(X=0)=\frac{\dfrac{1!}{0!\left(1-0\right)!}\ \dfrac{\left(1095-1\right)!}{\left(36-0\right)!\left(1095-1-36+0\right)!}}{\dfrac{1095!}{36!\left(1095-36\right)!}}=0.968$$

$$\Pr(X=1)=\frac{\dfrac{1!}{1!\left(1-1\right)!}\ \dfrac{\left(1095-1\right)!}{\left(36-1\right)!\left(1095-1-36+1\right)!}}{\dfrac{1095!}{36!\left(1095-36\right)!}}=0.032$$

$$\Pr(X\le 1)=0.968+0.032=1.000\Rightarrow \Pr(X\ge 2)=1-(0.968+0.032)=0.0$$

Therefore, the 1-sided upper-tailed p-value is zero. This is intuitive; if the target population truly contains only 1 exceedant day, then the probability that there will be more than one exceedance in any sample drawn from it will be zero. Thus if we find even 1 exceedance in a sample of 36, we must reject the null hypothesis that there was only 1 exceedance in the target population during the 3-year monitoring period. This result is more understandable if we think in terms of the proportion of exceedances. A sample with 1/36 exceedant days implies a population with an exceedance rate of 0.0278. Since there are 1095 days in the population, the expected number of exceedant days is 0.0278×1095˜ =31 days.

Using Equations 5 and 6, we can obtain confidence intervals for the total number exceedant of days. In this case we would like to compare the lower 1-sided 95% confidence limit to the hypothesized value of 1 exceedant day, so we use Eq. 5 and find that the lower bound is 2 exceedant days. This is greater than the hypothesized value (1) so it agrees with the exact p-value; however, we should note that the lower one-sided confidence interval on the estimate is extremely small relative to the estimate itself (i.e., 31), suggesting that the sample size of n=36 yields very low power. We can further illustrate this problem by comparing the probabilities of obtaining 0 exceedance from a population with only 1 exceedance (p=0.968) to the probability of obtaining 0 exceedances for a population with 20 exceedances (i.e., k=20). We obtain the later probability as,

$$\Pr(X=0)=\frac{\dfrac{20!}{0!\left(20-0\right)!}\ \dfrac{\left(1095-20\right)!}{\left(36-1\right)!\left(1095-20-36+1\right)!}}{\dfrac{1095!}{36!\left(1095-36\right)!}}=0.509$$

Thus, even if the true population has as many as 20 exceedant days there is still a 50-50 chance of obtaining 0 exceedant days in a sample of 36 days. This means that finding 0/36 exceedances does not provide a great deal of assurance that the true population does not contain substantially more than 1 exceedant day; i.e., the false negative error rate with n=36 is quite high (0.935) and the statistical power is low (0.065). A sample with zero out of 36 exceedances will not provide reliable evidence for attainment of the CMC standard.

Because the once-in-3-years criterion specifies that 2 or more exceedances in a year is a violation of the standard, if one assumes that there is no measurement error in the determination of the concentration of the pollutant, then any sample with 2 or more exceedant days will automatically indicate (i.e., without the need for statistical testing or estimation) that the target population had at least 2 exceedant days during the 3-year monitoring period. Thus the only situations where hypothesis testing and/or confidence interval estimation will be necessary is when there are only 0 or 1 exceedant days in the sample. Due to relationships described above, the occurrence of 1 exceedant day in a sample of n = 245 days will indicate rejection of the null hypothesis at the a=0.05 level.

Hence for any feasible sample size (e.g., 36 monthly assessments), the occurrence of even one exceedance will lead to a decision to list the body of water. This may appear to be excessively conservative at first; however, given sample sizes on the order of 36, most bodies of water with 1 or even 2 exceedances in a 3-year period *will not be listed*. This is because for a body of water that actually experiences 2 days during which a pollutant concentration exceeds the criterion, the probability that a sample of 36 assessments will contain at least one of the two exceedances is only 6.4%. Consequently a sample size of 36 will be associated with a false negative rate (i.e., Type II error rate) of 93.6%. Moreover, the probability that a sample of 36 assessments will contain both exceedances is only 0.1%.

The problem of determining the minimum sample size required for exact hypergeometric tests that maintain simultaneously specified false positive and false negative error rates is tedious without the aide of statistical software. At present, only one readily available statistical software package (PASS) will compute power and sample size for the exact hypergeometric distribution. Thus it is more difficult to apply the DQO procedure to the once-in-3-years CMC criterion based on the exact hypergeometric estimates, than for some of the other distributions discussed in this appendix (e.g., chi-square or t).

For either the acute or the chronic criteria, the Type I (false positive) and Type II (false negative) error rates and the power of the exact hypergeometric test against the upper 1-sided alternative for samples that contain 1 exceedant day can be specified as follows:

First, find the largest value of x_0 (i.e., the largest number of exceedant days) such that:

$$\Pr\left[x \geq x_0 \middle| k, n, N\right] \leq a \tag{7}$$

That is, we specify a critical value x_0 for the number of exceedant days, so that given the hypothesized number of exceedant days in the 3-year period (k), our sample size, and the population size, the probability of observing a larger number of exceedant days, is less than or equal to the alpha-level. For the situation under consideration that number is 1 day.

Next, having specified a value for x_0, we calculate the power and Type II error rate as:

$$\begin{aligned} &Power\left|k = \Pr\left[x \geq x_0 \middle| n, k, N\right] = 1 - \Pr\left[x < x_0 \middle| n, k, N\right]\right. \\ &\Rightarrow False\ Negative\ Error\ Probability\left|k = \Pr\left[x < x_0 \middle| n, k, N\right]\right. \end{aligned} \tag{8}$$

Thus, for specified N and a population that is assumed to have the smallest number of exceedances that will lead to a non-attainment decision (i.e., k=2), we can find the minimum sample size as the smallest value of n, in Eq. 8, that will yield a Power = 1-ß (where ß = the largest acceptable false positive error rate specified in the DQO). For the selenite example, to achieve a power of at least 0.85, the minimum sample size is n=1010. By contrast, when a sample of only 36 days is used, the power is only 0.064.

The same approach can be used to make inferences about the total number of chronic assessments in a 3-year period. Of course, chronic assessments are based on 4-day means, so there are only 1095/4 = 274 possible chronic assessments that could be made in any one 3-year period. Computations for confidence intervals and exact p-values can be made using equations 3 –6, substituting 274 for all values of N. Employing Eq. 4 to obtain the CDF for the case of x=1 exceedant mean, we find that the p-value for the corresponding test against the upper one-sided alternative is 0.13, indicating support for H_0. The estimated total number of chronic exceedances in the 3-year population, based on a sample with 1 chronic exceedance out 36 monthly assessments, is 8 exceedant days with a lower 95% bound of 1 exceedant day. Since the lower bound is within the acceptance region (i.e., it is = 1 exceedant day), a sample of only one exceedance out of 36 sampling units offers support for the null hypothesis that the body of water attains the chronic CMC standard; however, the probability of a false negative decision error with n=36 and a population size of N=274 is 0.9832. Thus a sample of n=36 that has only 0 or 1 exceedant days from a population that actually has 2 exceedant days, is almost certain lead to an incorrect attainment decision; sample sizes this small will not meet the DQO requirement that false negative error rates be less than 0.15. In fact the minimum sample size required for a false negative error rate = 0.15 is n=253.

The preceding calculations and discussion assume that exceedances occur independently of one another during the 3-year assessment period. However, it is more likely that exceedances will be associated with transient episodic events of variable duration. For example, consider a single discharge of flyash that elevates selenite concentrations above the CMC for 20 consecutive days in a reach of river during a 3-year assessment period. If the reach were being monitored on a regular monthly basis, there would be a 20/30.4 (66%) probability that a single exceedance would be recorded, a 34% probability that no exceedances would be recorded, and a 0.0% probability that two or more exceedances would be recorded during the 20-day episode.

The difficulties associated with the once-in-3-years assessments occur because the regulation allows 1 extremely rare event (e.g., 1 exceedant day out of 1095 or 1 out of 274 days), but not 2 extremely rare events. Thus, the width of the gray region is only 1/1095 for acute criteria or 1/274 for chronic criteria. When criteria are based on the assessment of such rare events, the false negative decision error rates become inflated to very high levels unless samples sizes are increased so that n ˜ N.

D.7 References

Altman, M.J., D. Machin, T. Bryant, and M. Gardner. 2000. Statistics with Confidence (2nd Ed.). BMJ Books, London, UK.

Beyer, W.H. (Ed.). 1971. CRC Basic Statistical Tables. CRC, Cleveland, Ohio.

Buonaccorsi, J. 1987. A note on confidence intervals for proportions in finite populations. American Statistician 41(3):215-218.

Brockwell, P.J. and R. A. Davis. 1987. Time Series: Theory and Methods. Springer-Verlag. New York.

Cleveland, W.S. 1993. Visualizing Data. Hobart Press. Summit, New Jersey, USA.

Cochran, W.G. 1977. Sampling Techniques (3rd ed.). New York, Wiley.

Diggle, Peter. 1990. Time series: a biostatistical introduction. Oxford University Press. Oxford, UK.

Freedman, D., R. Pisani and R. Purves. 1997. Statistics 3rd Ed. W.W. Norton & Co. New York

Hollander, M. and D. A. Wolfe. 1999. Nonparametric Statistical Methods, 2nd Ed. J. Wiley & Sons. New York.

Johnson, N. L. and S. Kotz.. 1969. Discrete Distributions. Houghton Mifflin. New York City, NY.

Lindley, D. and W. F. Scott. 1995. New Cambridge Statistical Tables (2nd Ed.) Cambridge University Press, New York. NY.

Millard, S. P. and N. K. Neerchal. 2000. Environmental Statistics with S-PLUS. CRC Press, Boca Raton, FL.

O'Brien, R.G. 1998. A tour of UnifyPow: a SAS module/macro for sample size analysis. Proceedings of the 23rd SUGI Conference, 1346-1355. SAS Institute, Cary, NC.

Smith, E. P., K. Ye, C. Hughes, and L. Shabman. 2001. Statistical assessment of violations of water quality standards under Section 303(d) of the Clean Water Act. Environ. Sci. Technol. 35:606-612.

Steel, G.D., J.H. Torrie and D.A. Dickey. 1996. Principles and procedures of statistics: a biometrical approach. McGraw-Hill, New York.

Stephan, C.E., D.I. Mount, D.J. Hansen, J.H. Gentile, G.A. Chapman, and W.A. Brungs. 1985. Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and their Uses: EPA.

Thompson, S.K. 1992. Sampling. J. Wiley and Sons. New York, New York.

Wendell, J. and J. Schmee. 2001. Likelihood confidence intervals for proportions in finite populations. American Statistician 55(1):55-61.

D.8 Glossary

alternative hypothesis - In a statistical hypothesis test there are two competing hypothesis, one of which, the alternative hypothesis, describes the set of conditions complementary to those described under the null hypothesis. For example: if the null hypothesis states that the mean pH of a set of samples is less than or equal to 5.0, the alternative hypothesis must be that the mean pH is greater than 5.0.

ANOVA Model - an acronym for Analysis of Variance. ANOVA models are linear models in which the total variance in a response is partitioned into two components: one due to treatments (and possible interactions among them) and the other due to random variability. In the simplest case where there is only one treatment factor, if the treatments have no effect on the response, the ratio of the variance components should be close to 1.0. If the treatments effect the response means, the ratio of the treatment component to the random component will be greater than one. Under the null hypothesis that the treatments have no effect, the sampling distribution of the ratio of the two variance components, each divided by their respective **degrees of freedom**, will be an F-distribution.

ARIMA Model - an acronym for autoregressive integrated moving average model. ARIMA models are linear models for regression and/or discrete treatment effects, measured through time, on responses that have been differenced at the appropriate **lag** distances.

autocorrelation - the internal correlation of a set of measurements taken over time and/or space. The correlation arises from the fact that points closer together in space and/or time tend to be more alike than those that are further apart. The autocorrelation function (either spatial or temporal) is a mathematical expression that relates the strength of the correlation to the distance (called the **lag**) between measurements.

Bayesian statistical inference - An approach to **inference** or **estimation** in which a process (e.g., a random binomial process) is proposed for the generation of a set of data. A mathematical model called a **likelihood** is specified for the process, such that the model parameters are random variables. A distribution, called the prior distribution, is developed for the parameters based on what is known about them, prior to collection of the data. Data are then collected and a mathematical principle called Bayes theorem is used to derive a second distribution of the parameters, called the posterior distribution, from the data and the prior distribution. The appropriate inference is then obtained from the posterior distribution. The Bayesian approach differs from the the classical frequentist approach in that is utilizes the investigator's prior knowledge of the system through the prior distribution.

bias - the systematic or persistent distortion of a measurement process that causes errors in one direction.

binary characteristic - a characteristic that can only have two possible values.

census - a study that involves the observation and/or measurement of every member of a population.

confidence interval - a range of values, calculated from the sample observations, that is believed, with a particular probability, to contain the true population parameter value. For example, a 95% confidence interval implies that if the estimation process were repeated again and again, then 95% of the calculated intervals would be expected to contain the true parameter value.

confidence level (also called the the confidence coefficient) - the probability that the confidence interval will include the true parameter value; Equivalently, 1-the probability (α)that the true value is *not* contained within the interval .

continuous random variable - A random variable which may take on an infinite number of values.

convenience sample - a sample collected from a target population without implementation of a probability-based design. Sampling units are selected based on ease of data collection, without clear reference to an underlying frame; e.g., the collection of water samples near bridges rather than randomly throughout the stream reach to which an inference is desired. Because many (perhaps the majority) of the population sampling units have little or no probability of selection to the sample and because sample coverage typically is restricted to some potentially homogeneous subset of the target population, data from convenience samples are not valid for statistical inference to the target population.

correlation coefficient – A scale-invariant measure of the association between 2 variables that takes on values between –1.0 and +1.0. The correlation coefficient has a value of plus one whenever an increase in the value of one variable is accompanied by an increase in the other, zero when there is no relationship (i.e., the 2 variables are independent of one another), and minus one (-1) when there is an exact inverse relationship between them.

correlogram - a plot or graph of the sample values of the autocorrelation coefficient of a time series against different values of the **lag**.

decision error - an error that occurs when data misleads an investigator into choosing the wrong response action, in the sense that a different action would have been taken if the investigator had access to unlimited "perfect data" or absolute truth. In a statistical test, decision errors are labeled as false rejection (Type I) or false acceptance (Type II) of a null hypothesis.

degrees of freedom (df) - As used in statistics, df has several interpretations. A sample of *n* values is said to have *n* degrees of freedom, but if *k* functions of the sample values are held constant, the number of degrees of freedom is reduced by *k*. In this case, the number of degrees of freedom is conceptually the number of independent observations in the sample, given that *k* functions are held constant. By extension, the distribution of a statistic based on *n* independent observations is said to have *n-p* degrees of freedom, where p is the number of parameters of the distribution.

discrete random variable - A random variable which may take on only a finite number of values.

dispersion - the amount by which a set of observations are spread out from their mean and/or median.

effect size - In a **one-sample test**, the difference between the sample mean and a pre-specified criterion or standard value. In a **two-sample test**, the effect size is the expected difference between the mean of a treatment group or ambient site vs. the mean of a control group or reference site. Associated statistical tests typically evaluate the null hypothesis of a zero effect size vs. the alternative that the effect size is nonzero.

effective sample size - When data are collected from cluster-correlated populations, there is redundancy in the information carried by more highly correlated individuals. Thus, correlated individuals carry less information than do uncorrelated individuals. The effective sample size is the number of uncorrelated individuals from a simple random sample that would carry information equivalent to the information in the sample of correlated individuals. The effective sample size is always less than the apparent sample size; how much less, is a function of the strength of the correlation and the sampling design that was used to collect the data.

estimation - the process of providing a numerical value for a population parameter on the basis of information collected from a sample.

experimental design - the arrangement or set of instructions used to randomize subjects to specific treatment or control groups in an experimental study. Such a procedure generally insures that results are not confounded with other factors and thus provides scientifically defensible inferences regarding causal effects of the treatments.

exploratory data analysis (EDA) - an approach to data analysis that may reveal structure and/or relationships among measured or observed variables in a data set. EDA emphasizes informal graphical procedures that typically are not based on prior assumptions about the structure of the data or on formal models.

extreme values - the largest and smallest values (and perhaps their neighboring values) among a sample of observations.

frequentist statistical inference - an approach to statistics based on the likelihood of an observed result in a large or infinite number of independent repetitions of the same sampling or experimental procedure (e.g., see the frequentist definition of the **confidence interval** in this glossary).

geometric mean - a measure of central tendency calculated by back-transforming the mean of a set of log-transformed observations. If the original data come from a log-normal distribution, the sample geometric mean will provide an unbiased estimate of the sample **median**.

heterogeneous - a term denoting inequality or dissimilarity of some quantity of interest (most often a variance) in a number of different groups, populations, etc.

homogeneous - a term denoting equality or similarity of some quantity of interest (most often a variance) in a number of different groups, populations, etc.

imprecision/precision - A term describing the degree of spread among successive estimates of a population parameter by a sample statistic. The standard error of a sample estimator (e.g., the standard error of the mean) is a measure of imprecision/precision in the estimator. A high degree of spread (imprecision) will lead to an increased likelihood of a decision error, while a reduction in spread will lead to a corresponding reduction in the likelihood of a decision error. Generally, precision will be increased by increasing the sample size.

independence - essentially, two events are said to be independent if knowing the outcome of one tells us nothing about the outcome of the other. More formally, two events *A* and *B* are said to be independent if Probability (A and B) = Prob(A) × Prob(B).

inference - the process of drawing conclusions about a population on the basis of measurements or observations made on a sample of individuals from the population.

lag - the distance, in units of time or space, between two events or locations. For example, an event occurring at time $t+k$ $(k>0)$ is said to lag behind the event occurring at time t, by an amount of time equal to lag k.

likelihood - the probability of a set of observed data, given the value of some parameter or set of parameters associated with a model for the underlying process that is hypnotized to have generated the data. For example, if we obtain 9 heads in 10 tosses of a coin, the likelihood of observing this result, given that the coin is fair (i.e., the binomial parameter p=0.50), is approximately 0.0098.

log-transformation - a transformation on a variable, X, obtained as, Y=ln(X) or Y=ln(x+c), where c is a constant positive value (e.g., 1.0). This transformation is useful for normalizing continuous variables with skewed distributions and/or stabilizing the variance of a variable whose standard deviation increases as a function of its mean, prior to statistical analysis with tests (e.g., t-tests) that assume normality and/or variance homogeneity..

maximum likelihood - a procedure for estimating the value of a parameter(s) of a model of the underlying process that produced some particular set of observations, such that the resulting estimate maximizes the likelihood of the observed data. For example, the maximum likelihood estimate for the binomial parameter P, given an experiment in which one obtains 9 heads in 10 tosses is P=0.90. The likelihood of obtaining 9 heads given an underlying binomial process with P=0.90, is 0.3874. Note that the estimate P=0.90 leads to a much larger likelihood than an estimate of P=0.50 does (0.0098; see definition of **likelihood**). In fact there is no value of P that will yield a larger likelihood of obtaining 9 heads out of 10 tosses than the estimate P=0.90; thus, P=0.90 is the maximum likelihood estimator of P.

median - in a sample or a population, the median is the value of a random variable such that half of the sampling units have larger values and half have smaller values. When the population or sample size is 2N+1, the median is the value of the random variable associated with the N+1th ordered sampling unit; when the population or sample size is 2N, the median is average of random variable values of the sampling units with ranks N and N+1. If the population is normal the median will equal the mean. If the population is log-normal the median will equal the **geometric mean**.

Monte Carlo methods - methods for finding solutions to mathematical and statistical problems by simulation. Often used when the analytic solution of the problem is intractable, or when real data are difficult to obtain, or to evaluate the behavior of statistics or models under a variety of hypothetical conditions which may or may not be directly observable in nature.

nonparametric statistical methods (also called distribution-free methods) - Statistical techniques of **estimation** and **inference** are often based on the assumption of some underlying parametric process; for example, one that generates responses that are normally distributed. By contrast, nonparametric estimation and testing procedures do not depend on the existence of an underlying parametric process. Consequently, nonparametric techniques are valid under relatively general assumptions about the underlying population. Often such methods involve only the ranks of the observations rather than the observations themselves.

noncentral t-distribution - the expected distribution of the t statistic when the alternative hypothesis is true. This contrasts with central t-distribution (usually referred to simply as the "t-distribution") which is the expected distribution of the t-statistic when the null hypothesis is true. In general, the probability that an observed t-statistic comes from a non-central t-distribution will be large (e.g., P>0.20) when the probability of that it comes from a central t-distribution is low (e.g., P<0.001), and vice versa.

null hypothesis - a hypthesis about some presumed prevailing condition, usually associated with a statement of "no difference" or "no association" (see also **alternative hypothesis**).

one-sample tests - Statistical tests that evaluate the null hypothesis that there is no difference between a sample statistic (e.g., mean or proportion) and a fixed criterion or standard value.

parametric continuous distribution - the probability distribution of a continuous random variable, specified by a mathematical function of the population parameters; e.g., the normal distribution with parameters, μ and σ^2.

parametric statistical methods - tests and estimation procedures that depend on the complete specification of an underlying parametric probability distribution of the population from which the sample was drawn. The estimators and test statistics that are based on functions of the estimates of the population parameters under the assumed population distribution model (e.g. normal) are valid only if the assumed population model is valid. An example is the t-statistic which assumes an underlying normal population.

percentiles - the set of divisions of a set of data that produce exactly 100 equal parts in a series of values.

population - any finite or infinite collection of "units" that completely encompasses the set individuals of interest. In environmental studies, populations are usually bounded in space and time; e.g., the population of smallmouth bass in Leech Lake, Minnesota on July 1, 2000.

population parameter - a constant term(s) in a mathematic expression, such as a probability density function, that specifies the distribution of individual values in the population. Parameters typically control the location of the center of the distribution (location parameters), the spread of the distribution (scale or dispersion parameters) and various aspects of the shape (shape parameters) of the distribution (see also: **probability density function**).

power of a statistical test -the probability of rejecting the null hypothesis when it is false. Notice that we would like always to reject a false hypothesis; thus, statistical tests with high power (i.e., power >0.80) are desirable. Generally the power of a test increases with the number of individuals in the sample from which the test was computed.

precision - a term applied to the uncertainty in the estimate of a parameter. Measures of the precision of an estimate include its standard error and the confidence interval. Decreasing the value of either leads to increased precision of the estimator.

probability-based sample - a sample selected in such a manner that the probability of being included in the sample is known for every unit on the sampling frame. Strictly speaking, formal statistical inference is valid only for data that were collected in a probability sample.

probability density function (PDF)- for a continuous variable, a curve described by a mathematical formula which specifies, by way of areas under the curve, the probability that a variable falls within a particular range of values. For example, the normal probability density function of the continuous random variable X, is:

$$\frac{1}{s\sqrt{2p}}\exp\left[-\left(\frac{1}{2s^{2}}\right)(x-m)^{2}\right]$$

The normal probability density function has two **parameters**, the mean and variance , μ and σ^2. The mean is the location parameter and the variance is the scale parameter; the normal distribution does not have any shape parameters. The graph of the normal probability density function is the familiar "bell curve".

rank - the relative position of a sample value within a sample.

relative frequency - the frequency of occurrence of a given type of individual or member of a group, expressed as a proportion of the total number of individuals in the population or sample that contains the groups. For example, the relative frequencies of 14 bass, 6 bluegill, and 10 catfish in a sample of 30 fish are, respectively: 46.7%, 20.0% and 33.3%.

representative sample - A sample which captures the essence of the population from which it was drawn; one which is typical with respect to the characteristics of interest, regardless of the manner in which it was chosen. While representativeness in this sense cannot be completely assured, probability-based samples are more likely to be representative than are judgement or convenience samples. This is true because only in probability sampling will every population element have a known probability of selection.

sample - a set of units or elements selected from a larger population, typically to be used for making inferences regarding that population.

sampling design - a protocol for the collection of samples from a population, wherein the number, type, location (spatial or temporal) and manner of selection of the units to be measured is specified.

sampling distribution - the expected probability distribution of the values of a statistic that have been calculated from a large number of random samples. For example, the sampling distribution of the ratios of each of the means from 100 samples (each with n=30) to their respective variances will be a t-distribution with 29 degrees of freedom.

sampling error - the difference between a sample estimate and the true population parameter due to random variability in the composition of the sample vs. that of the target population.

sampling frame - the list from which a sample of units or elements is selected.

sampling unit - the members of a population that may be selected for sampling.

significance level (α) - the level of probability at which it is agreed that the null hypothesis will be rejected; α is also the probability of a Type I error.

skewness - a measure of the asymmetry in a distribution, relative to its mean. A right-skewed distribution is composed mostly of small values lying close to the mean but possesses a few values that are much larger than the mean. Conversely, a left-skewed distribution is composed mostly of values lying close to the mean but possesses a few values that are much smaller than the mean.

square-root transformation - a transformation on a variable, X, obtained as, Y= or Y=$\sqrt{X + \frac{1}{2}}$. This transformation is useful for normalizing a discrete variable with a Poisson distribution and/or stabilizing the variance of a variable whose variance is proportional to its mean, prior to statistical analysis with tests (e.g., t-tests) that assume normality and/or variance homogeneity.

standard deviation - the square root of the **variance**.

standard error - the standard error of a sample statistic, θ, (say a sample mean or proportion) is the standard deviation of the values of that statistic computed from repeated sampling of the target population, using the same sampling design (e.g., stratified simple random sampling) and the same sample size, *n*. For example, the standard error of the mean is the sample **standard deviation**/n.

standard normal distribution - a normal distribution whose mean is 0 and whose variance is 1.

statistic - a quantity calculated from the values in a sample (e.g., the sample mean or sample variance)**.**

statistical distribution - a probability distribution used to describe a statistic, a set of observations or a population.

statistical test of hypotheses - a statistical procedure for determining if a sample provides sufficient evidence to reject one statement regarding the population of interest (the null hypothesis) in favor of an alternative statement (the **alternative hypothesis**).

target population - the set of all units or elements, bounded in space and time, about which a sample is intended to produce inferences.

two-sample tests - Statistical tests that evaluate the null hypothesis that there is no difference between a sample statistic (e.g., mean or proportion) in a treatment group or at an ambient monitoring site and the sample statistic in a control group or at a reference site.

Type I error (α) - the error that occurs when a decision maker rejects the null hypothesis when it is actually true. Also called the false rejection decision error, or false positive decision error.

Type II error (β) - the error that occurs when a decision maker accepts a null hypothesis when it is actually false. This is also called the false acceptance decision error, or false negative decision error. The power of a statistical test is 1-β.

variance (population) - the variance of a finite population of N values - $x_1,x_2,....x_N$ - is simply the average of the squared difference between the individual observations and the population mean.

variance (sample) - the variance of n sample observations is simply the average of the squared differences between the individual observations and the sample mean, divided by (n-1).

variogram - a plot of the sample values of the variance of a spatially referenced variable vs. the corresponding lag distances